IN THE UNITED STATES DISTRICT COURT
FOR THE WESTERN DISTRICT OF TEXAS
WACO DIVISION

XOCKETS, INC., Plaintiff,	
v.	Civil Action No. 6:24-cv-453
NVIDIA CORPORATION, **MICROSOFT CORPORATION, and** **RPX CORPORATION,**	**JURY TRIAL DEMANDED**
Defendants.	

COMPLAINT FOR VIOLATION OF FEDERAL ANTITRUST LAWS AND REQUEST FOR INJUNCTION

TABLE OF CONTENTS

Plaintiff Xockets, Inc. ("Plaintiff" or "Xockets") hereby submits this Complaint against Defendants NVIDIA Corporation ("NVIDIA"), Microsoft Corporation ("Microsoft"), and RPX Corporation ("RPX") for violation of federal antitrust laws for which it seeks an injunction, and states as follows:

1. NVIDIA holds monopoly power in the market for GPU-enabled artificial intelligence computer systems, holding market share above 90% by unit. Microsoft has combined its leading position in cloud services with control over leading generative AI models to maintain and/or create a monopoly in GPU-enabled generative artificial intelligence platforms. Microsoft has entered into unlawful agreements in restraint of trade with the owner of leading generative AI models, including OpenAI, and with the dominant supply of GPU-enabled artificial intelligence computer systems. NVIDIA and Microsoft have formed a cartel to create and/or maintain a monopoly in GPU-enabled generative artificial intelligence. As part of this cartel, NVIDIA and Microsoft have formed a buyers' cartel to avoid paying the fair market price for the fundamental intellectual property that transformed GPU and GPU-enabled platforms from a niche product for gamers and cryptocurrency miners into the most important industrial component in the United States economy today. The fundamental technology was created by Xockets. This buyers' cartel is designed to fix below market level the price for the critical technology held by Xockets.

2. The buyers' cartel at issue in this case is part of a pattern of illegal cartel behavior engaged in by NVIDIA and Microsoft, as evidenced by the ongoing investigations of these entities by the United States Department of Justice, the United States Federal Trade Commission, and the European Union.

3. The illegal buyers' cartel in this case was facilitated by RPX. RPX was formed at the request of Big Tech companies to enable and create buyers' cartels for intellectual property.

RPX previously touted on its website that "**[i]n effect, RPX can buy 'wholesale' on behalf of our client network, while our clients otherwise would pay 'retail' if transacting on their own.**" In subsequent years RPX tried to erase this admission from the web because the "client network" as RPX euphemistically describes it consists of the companies that hold monopoly power in essentially every high technology field in this Country. This includes NVIDIA and Microsoft, who engaged RPX to manage the buyers' cartel for the fundamental intellectual property that drives the AI revolution.

4. The fundamental intellectual property that turned NVIDIA's GPUs from niche equipment for gamers and cryptocurrency miners to the driver of the AI revolution was created by Xockets. Every generation NVIDIA makes incremental improvements in its GPUs (including its Hopper and upcoming Blackwell GPUs) through a combination of leveraging improvements in manufacturing technology of the companies which manufacture its chips and minor improvements in its architecture.

5. What has allowed NVIDIA to monopolize the field of GPU-enabled artificial intelligence servers that third parties can buy is the introduction of three other entirely different components that use Xockets' patented technology. These components, or Data Processing Units (DPUs), include the BlueField, ConnectX, and NVLink Switch DPUs for offloading, accelerating, and isolating data-intensive workloads from server processors in cloud data centers. They are necessary for allowing NVIDIA to combine a large number of GPU-enabled server boards or modules in order to process the vast amounts of data necessary for artificial intelligence and to provide this technology to customers across the web. NVIDIA's CEO described the importance of its DPUs:

> "The holy trinity of computing … is the CPU, the GPU, and the DPU. These three processors are fundamental to computing."[1]
>
> "The data center has become the new unit of computing. DPUs are an essential element of modern and secure accelerated data centers in which CPUs, GPUs and DPUs are able to combine into a single computing unit that's fully programmable, AI-enabled and can deliver levels of security and compute power not previously possible."[2]
>
> "A single BlueField-2 DPU can deliver the same data center services that could consume up to 125 CPU cores."[3]

6. NVIDIA's CEO also described the critical importance of its NVLink Switch DPUs: "for large language models like the Chat GPT and others like it . . . all these GPUs have to share the results, partial products [of AI model training operations] . . . Whenever they do all-to-all, all-gather, whenever they communicate with each other, that NVLink Switch is communicating almost 10 times faster than what we could do in the past using the fastest networks."[4] "The miracle is this chip—this NVLink Chip."[5]

7. Since the introduction of Xockets' patented designs for its DPUs, NVIDIA's market capitalization has exploded, from $180 billion to approximately $3 trillion.

8. NVIDIA did not invent the technology in its BlueField, ConnectX, and NVLink Switch DPUs. This technology was taken from Xockets. And it was done so knowingly. Instead

[1] https://www.nextplatform.com/2020/04/27/nvidia-plus-mellanox-talking-datacenter-architecture-with-jensen-huang.

[2] https://nvidianews.nvidia.com/news/nvidia-introduces-new-family-of-bluefield-dpus-to-bring-breakthrough-networking-storage-and-security-performance-to-every-data-center.

[3] https://nvidianews.nvidia.com/news/nvidia-introduces-new-family-of-bluefield-dpus-to-bring-breakthrough-networking-storage-and-security-performance-to-every-data-center.

[4] GTC March 2024 Keynote with NVIDIA CEO Jensen Huang, https://www.youtube.com/watch?v=Y2F8yisiS6E&t=3403s (56:43–59:06).

[5] NVIDIA CEO Jensen Huang Keynote at COMPUTEX 2024, https://www.youtube.com/watch?v=pKXDVsWZmUU&t=4283s (1:11:23–1:15:57).

of paying fair value for the technology, Microsoft and NVIDIA took it without permission. And when NVIDIA and Microsoft were put on notice that Xockets would not allow its technology to be used illegally without a license they formed an illegal buyers' cartel with RPX to drive the price of Xockets' patents on its DPU inventions below the market price and/or drive Xockets out of business. This lawsuit will stop the illegal conduct.

INTRODUCTION

I. XOCKETS AND ITS DPU INVENTIONS

9. Xockets was founded in 2012 by Dr. Parin Dalal and a team of network infrastructure engineers, turned early cloud engineers. Dr. Dalal received a bachelor's degree in computer science from University of California, Berkeley and his Ph.D. in theoretical physics from the University of California, San Diego, and began his career as an engineer designing CPUs and GPUs. Today, Dr. Dalal is Principal Engineer, Machine Learning and Artificial Intelligence, at Google. Prior to joining Google Dr. Dalal led company-wide strategic AI decision-making at Varian Medical Systems, now a Siemens company, as its Vice President of Advanced AI developing AI-based formulations of cancer treatments to save lives.

10. Founding investors in Xockets include the current CTO of Intel, Dr. Greg Lavender, who was a long-time Xockets Board member and considered Xockets' DPU architecture a "transformational" and "revolutionary" new computing architecture in clouds; Robert Cote, one of the nation's leading IP investors and lawyers, who guided the company in ensuring that Xockets' breakthrough DPU inventions were protected by United States patents; Jerry Yang, cofounder of Yahoo, who invested through AME cloud ventures, a venture capital firm he founded to invest in breakthrough cloud technologies.

11. In the early 2010s, Xockets' co-founder and lead inventor, Dr. Parin Dalal, had the vision to see that conventional wisdom in the computing industry—that relies on expected

increases in transistor density for ever-faster computing performance known as Moore's Law—would fail to meet the unique challenges that data-intensive workloads would pose to distributed computing performance in cloud data centers. He foresaw that Moore's Law would end as the data workloads driving distributed computing in clouds would grow by orders of magnitude. The manipulation and analysis of vast data sets across GPU-enabled server processors in the training of large language models, like ChatGPT, requires that this problem be addressed to enable the age of artificial intelligence that is now underway in the world.

12. Dr. Dalal realized that a new computing paradigm, involving a new accelerated computing architecture that extends into the network of cloud data centers, was needed. This new computing architecture for processing data-intensive workloads was implemented by Xockets in a new cloud processor known today as a Data Processing Unit, or DPU. It is designed to provide flexible hardware-like handling of computing operations at the speed of the network—or line rate—with software-like programmability that can form programmable logic pipelines of hardware accelerators for processing data-intensive workloads independent of server processors and conventional computing architectures. This programmable hardware acceleration in the network invented by Dr. Dalal can run new, varied, and evolving cloud infrastructure services. It provides the versatility clouds require to offload infrastructure services and accelerate many different kinds of data-intensive workloads and processes, freeing up server processors to run their main workloads or applications for customers at ever-increasing speeds and lower power costs.

13. Xockets described Dr. Dalal's inventions in a series of patent applications filed with the United States Patent Office beginning in May 2012, and did so in reliance on the promise made in the United States Constitution that Xockets would be granted exclusive rights to Dr. Dalal's DPU inventions. These exclusive rights were placed by Congress in the United States Patent Laws

and are fundamental to the innovation economy that our nation's Founders sought to build in a country that was once a startup nation. With this promise, people from all over the world and from all walks of life came to this country and created an innovation economy that is unparalleled in history—from which was built the largest economy and most prosperous nation on earth.

14. The growth of this innovation economy depends on the strict enforcement of an inventor's exclusive rights as promised in the United States Constitution. Indeed, the most important inventions in our nation's history have come from entrepreneurs like Dr. Dalal and startups like Xockets, not from those who hold the reins of power. The strict enforcement of intellectual property rights is what creates a level playing field for inventors and incentivizes the personal sacrifice that these entrepreneurs must make to build a future to benefit us all, and it is what instills confidence in investors to invest in breakthrough new startups like Xockets and inventors like Dr. Dalal.

15. To date, Xockets has obtained a number of patents covering many aspects of Dr. Dalal's DPU inventions—as there were many problems to solve. Xockets currently has over 60 patent applications prepared for filing directed to numerous other DPU inventions. Xockets' issued patents include: (i) Xockets' DPU Computing Architecture Patents (also known as the "***New Cloud Processor Patents***"), including U.S. Patent Nos. 11,080,209 ("the '209 Patent" – DPU Computing Architecture, Security), U.S. Patent No. 10,649,924 ("the '924 Patent" – DPU Network Overlay, Security), and U.S. Patent No. 11,082,350 ("the '350 Patent" – DPU Stream Processing); and (ii) Xockets' DPU Switching Architecture Patents (also known as the "***New Cloud Fabric Patents***"), including U.S. Patent No. 10,223,297 ("the '297 Patent" – DPU Cloud Network Fabric), U.S. Patent No. 9,378,161 ("the '161 Patent" – DPU Cloud Network Fabric), U.S. Patent No. 10,212,092 ("the '092 Patent" – DPU In-Network Computing), and U.S. Patent No. 9,436,640

(“the ’640 Patent” – DPU In-Network Computing). These patents, including the ’209 Patent, ’924 Patent, ’350 Patent, ’297 Patent, ’161 Patent, ’092 Patent, and ’640 Patent, are collectively referred to herein as the “Asserted Patents” or “Xockets Patents.” In addition to being infringed, the Xockets Patents are targets of Defendants’ unlawful buyers’ cartel.

16. Xockets’ patented inventions include a groundbreaking new cloud computing architecture and a new cloud network fabric—a reinvention of cloud distributed computing from the ground up—that serve to dramatically increase the speed and lower the costs of distributed computing services and AI. To do so, Xockets’ patented DPU architecture enables cloud server computers to offload, accelerate, and isolate critical data-intensive tasks that would otherwise overburden server processors.

17. Xockets’ New Cloud Processor Patents, for example, describe a virtual switch computing architecture for offloading from server processors to DPUs, or offload processor modules, accelerating, and isolating data-intensive workloads in clouds such as security, networking, and storage computing operations in moving data between server processors. In other words, Xockets’ Patents describe the use of a virtual switch computing architecture in a new cloud processor for brokering collective communication between server processors (the movement of data between CPUs, GPUs, and hybrids of these server processors) independent of the limitations of conventional computing architectures, and for freeing up server processors to run customer applications at higher speeds and lower cost.

18. Xockets’ New Cloud Fabric Patents, for example, further describe connecting together these DPUs in a novel way to form a new cloud network fabric for brokering collective communication independent of the limitations of existing cloud networks. This new cloud fabric is designed for even faster, lower-cost collective communication among server processors, and for

in-network computing operations such as sorting, organizing, and reducing/combining data-intensive workloads in distributed computing. This new cloud fabric enables the training of large AI models across GPUs in a matter of weeks or months rather than many years as would otherwise be required. In this way, Xockets' DPU inventions ensure that training large models for AI and the production of AI can be made widely available and affordable to every business in every industry to drive forward a new industrial revolution.

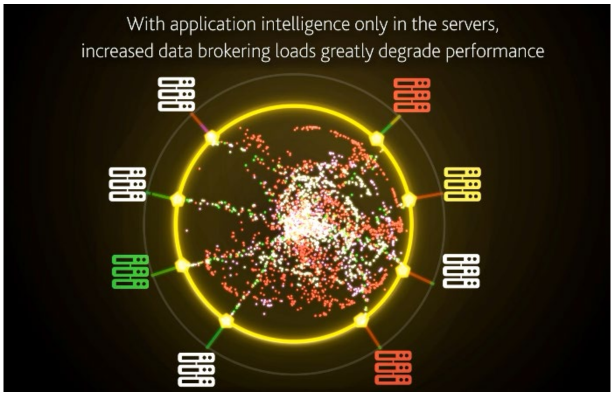
II. XOCKETS PRESENTED ITS PATENTED DPU TECHNOLOGY, WHICH WAS THEN STOLEN BY DEFENDANTS

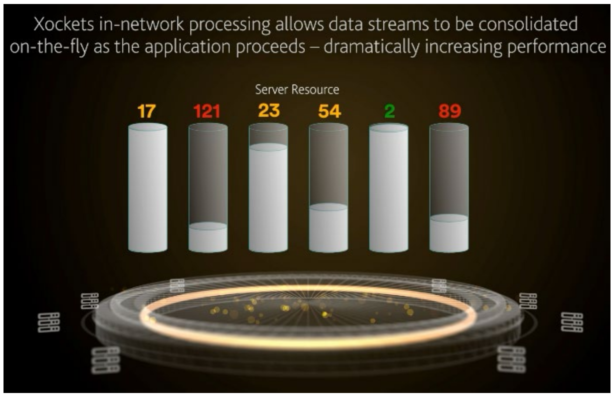
19. Xockets developed the world's first DPUs in a product it called the StreamSwitch. Xockets publicly displayed its patented DPU architecture in the StreamSwitch at Strata, the industry's premier big data and network technology conference, in the Fall of 2015.

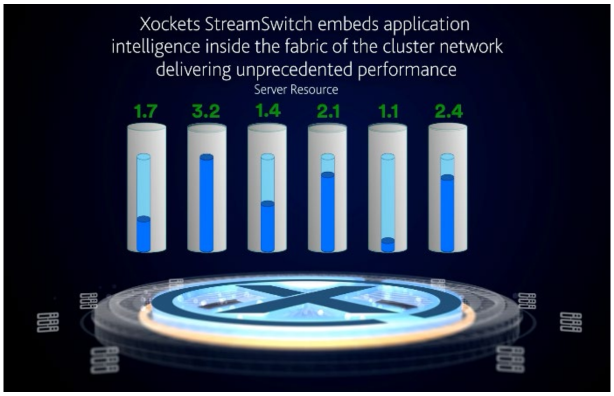
XOCKETS + SMART Modular Technologies
A Revolution in Scale-Out Architecture
StreamSwitch
Transforms Big Data Performance
Active Application Processing embedded into the network
Job Completion Time reduced 100X
No change to in-place hardware or software infrastructure
Reduces Big Data cluster TCO up to 40%
Active Application Network
Solve your hardest computational problems in real time with no change to your existing cluster and software tools
100X higher cluster efficiency for Hadoop
Delivering unprecedented perform
243mW / Gbps
StreamSwitch
XOCKETS + SMART Modular Technologies
A Revolution in Scale-Out Architecture

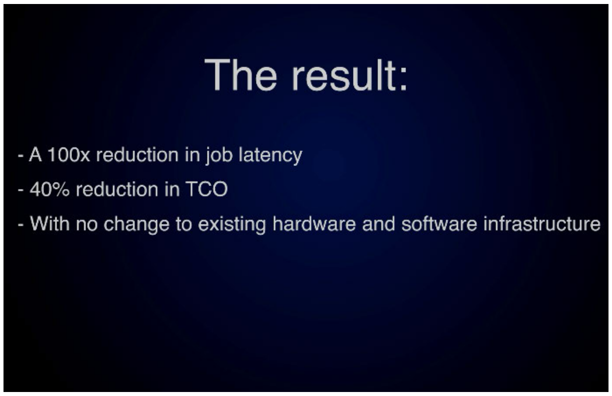
20. At the Conference, Xockets demonstrated its revolutionary new DPU computing architecture and the "unprecedented performance" benefits it provides by offloading, accelerating,

and isolating processing of data-intensive workloads from server processors and conventional computing, and by forming a new cloud network fabric of interconnected DPUs that can operate independent of the performance limitations of existing cloud networks:


With application intelligence only in the servers, increased data brokering loads greatly degrade performance

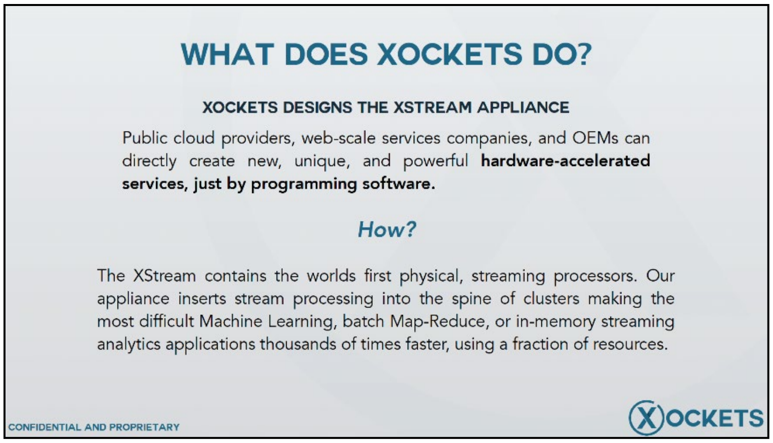

Xockets in-network processing allows data streams to be consolidated on-the-fly as the application proceeds – dramatically increasing performance


Xockets StreamSwitch embeds application intelligence inside the fabric of the cluster network delivering unprecedented performance


The result:

- A 100x reduction in job latency
- 40% reduction in TCO
- With no change to existing hardware and software infrastructure

21. Xockets presented its DPU technology to Microsoft in 2016, demonstrating the invention’s ability to accelerate computing performance on cloud data-intensive workloads—including “Big Data,” “Machine Learning” (which includes AI), “Security,” and “Encryption / Decryption” workloads—thousands of times faster using a fraction of the resources:



WHAT DOES XOCKETS DO?

XOCKETS DESIGNS THE XSTREAM APPLIANCE

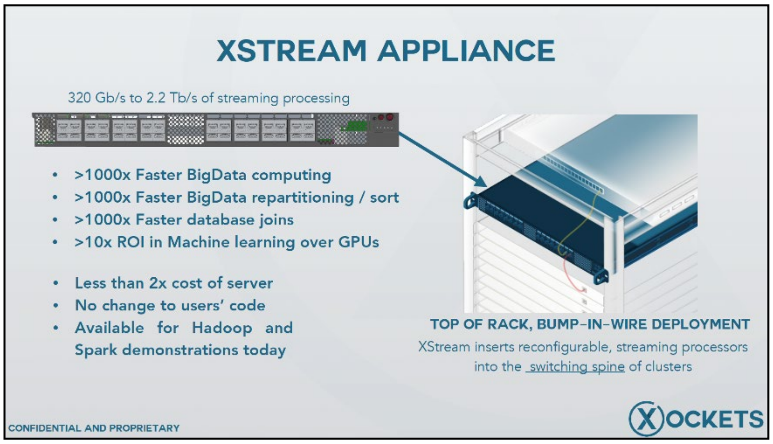
Public cloud providers, web-scale services companies, and OEMs can directly create new, unique, and powerful **hardware-accelerated services, just by programming software.**

How?

The XStream contains the worlds first physical, streaming processors. Our appliance inserts stream processing into the spine of clusters making the most difficult Machine Learning, batch Map-Reduce, or in-memory streaming analytics applications thousands of times faster, using a fraction of resources.

CONFIDENTIAL AND PROPRIETARY

XOCKETS



XSTREAM APPLIANCE

320 Gb/s to 2.2 Tb/s of streaming processing

- >1000x Faster BigData computing
- >1000x Faster BigData repartitioning / sort
- >1000x Faster database joins
- >10x ROI in Machine learning over GPUs

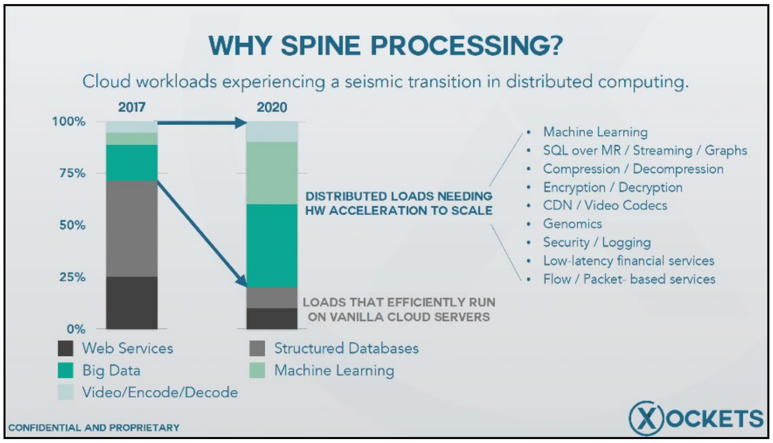
- Less than 2x cost of server
- No change to users' code
- Available for Hadoop and Spark demonstrations today

TOP OF RACK, BUMP–IN–WIRE DEPLOYMENT

XStream inserts reconfigurable, streaming processors into the switching spine of clusters

CONFIDENTIAL AND PROPRIETARY

XOCKETS



WHY SPINE PROCESSING?

Cloud workloads experiencing a seismic transition in distributed computing.

DISTRIBUTED LOADS NEEDING HW ACCELERATION TO SCALE

- Machine Learning
- SQL over MR / Streaming / Graphs
- Compression / Decompression
- Encryption / Decryption
- CDN / Video Codecs
- Genomics
- Security / Logging
- Low-latency financial services
- Flow / Packet- based services

LOADS THAT EFFICIENTLY RUN ON VANILLA CLOUD SERVERS

■ Web Services
■ Big Data
■ Video/Encode/Decode
■ Structured Databases
■ Machine Learning

CONFIDENTIAL AND PROPRIETARY

XOCKETS

22. Xockets' DPU technology was thereafter adopted by Mellanox in 2016[6] without Xockets' knowledge or permission for cloud offload use of server processors by Microsoft and other customers.

23. Mellanox's use of Xockets' DPU technology led to NVIDIA later acquiring Mellanox in order to drive forward NVIDIA's collaborations with Microsoft to dominate the markets for AI equipment and services using Xockets' DPU technology in its products. These collaborations continue to this day.

24. Thus, instead of licensing the technology, both Microsoft and NVIDIA chose to stand on the backs of Xockets' and Dr. Dalal's ingenuity, hard work, and innovations, and attempted to shut Xockets out of the market by misappropriating Xockets' technology and patents, all while NVIDIA was falsely proclaiming itself as the pioneer of accelerated computing and AI with its DPUs.

25. In addition, NVIDIA and Microsoft chose to form a cartel that leveraged Xockets' DPU technology to create the dominant market position these two companies hold today.

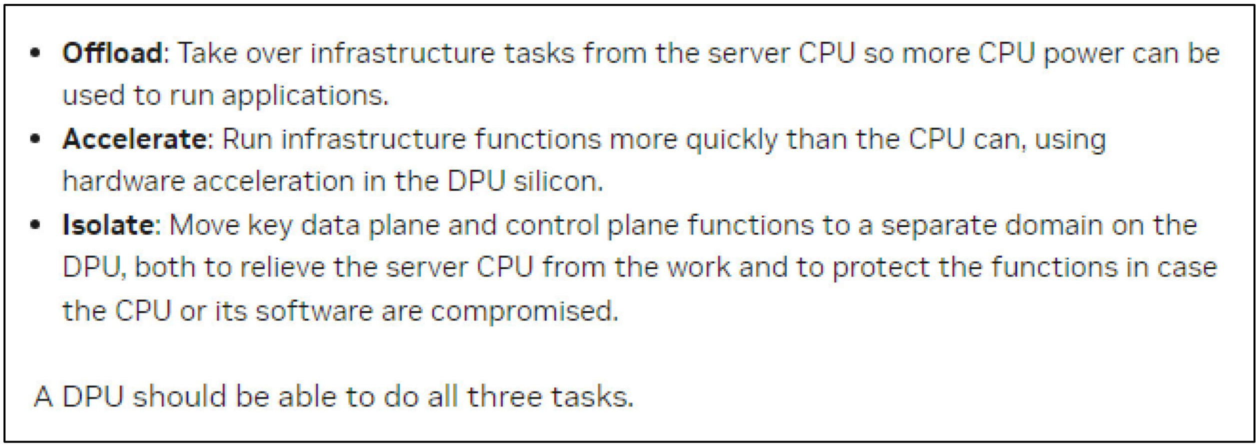
26. Microsoft has control over the leading generative artificial intelligence models in the world. Microsoft and NVIDIA have formed a cartel through an extensive series of what they euphemistically refer to as "collaborations" to maintain or create a monopoly in GPU-enabled generative artificial intelligence.

[6] https://www.businesswire.com/news/home/20160615005424/en/Mellanox-Announces-ConnectX-5-the-Next-Generation-of-100G-InfiniBand-and-Ethernet-Smart-Interconnect-Adapter; https://www.servethehome.com/mellanox-connectx-5-vpi-100gbe-and-edr-ib-review/mellanox-connectx-4-connectx-5-and-connectx-6-ethernet-comparison-chart-1.

III. XOCKETS' DPU INVENTIONS ARE ESSENTIAL ELEMENTS OF NVIDIA'S AND MICROSOFT'S GPU-ENABLED SERVER COMPUTER SYSTEMS

27. NVIDIA's systems feature distinct DPUs, including BlueField and ConnectX DPUs, as well as NVLink Switch DPUs, that offload key data-intensive workloads from server processors, including, among others, security, networking, and storage operations, and that form a virtual switching fabric for accelerating collective communication among server processors and enabling in-network computing of data-intensive workloads in training machine learning/artificial intelligence (ML/AI) models, all as claimed in the Xockets Patents.

28. For example, NVIDIA publicly states that "[t]he best definition of the DPU's mission is to offload, accelerate, and isolate infrastructure workloads" and further explains each function[7]:



- **Offload**: Take over infrastructure tasks from the server CPU so more CPU power can be used to run applications.
- **Accelerate**: Run infrastructure functions more quickly than the CPU can, using hardware acceleration in the DPU silicon.
- **Isolate**: Move key data plane and control plane functions to a separate domain on the DPU, both to relieve the server CPU from the work and to protect the functions in case the CPU or its software are compromised.

A DPU should be able to do all three tasks.

29. Microsoft is a customer of NVIDIA and with privileged access to NVIDIA's infringing GPU-enabled server computer systems and components for AI, which Microsoft uses in, inter alia, its Microsoft Azure Cloud computing platform.

[7] https://developer.nvidia.com/blog/offloading-and-isolating-data-center-workloads-with-bluefield-dpu.

30. Microsoft is combining its privileged access to NVIDIA's equipment and its control over the leading generative artificial intelligence models in the world to leverage NVIDIA's monopoly over AI equipment to create a monopoly in AI platforms based on the equipment. Microsoft and NVIDIA have formed a cartel to affect this process.

31. This case focuses on NVIDIA's and Microsoft's infringement of Xockets' inventions, including in GPU-enabled server computer systems that use Xockets' claimed DPU computing architecture (e.g., enabled by NVIDIA's BlueField and ConnectX DPUs) as well as Xockets' claimed DPU cloud network fabric architecture (e.g., enabled by NVIDIA's NVLink Switch DPUs). The infringing systems include, for example, NVIDIA's existing Hopper GPU-enabled server computer systems for AI and the greatly expanded infringement in NVIDIA's upcoming Blackwell GPU-enabled server computer systems for AI scheduled for release this Fall 2024.

32. NVIDIA has been making and selling DPUs for its cloud GPU-enabled server systems since at least as of April 2020, when it completed its purchase of Mellanox (for its Bluefield and ConnectX DPUs) for approximately $7 billion. NVIDIA's founder and CEO, Jensen Huang, declared NVIDIA's acquisition of Mellanox "a homerun deal":

> "This is a homerun deal. Man I've been dreaming about this. You know the most important computer today is the data center, it is the epicenter of the computer industry. And ***the most important applications that run in the data center today are AI applications and Big Data analytics applications. Doing computation on artificial intelligence . . . and moving huge amounts of data around is what drives the data center architectures today***. And so we are combining the leaders of AI computing and high speed networking and data processing into one company. This is really quite extraordinary."[8]

[8] https://www.cnbc.com/2020/04/27/nvidia-ceo-calls-mellanox-acquisition-a-homerun-deal.html. All emphases are added unless otherwise indicated.

33. Explaining the Mellanox acquisition, Huang also stated:

> "We believe that in future datacenters, the compute will not start and end at the server, but ***the compute will extend into the network*.** ***And the network itself, the fabric, will become part of the computing fabric.***"[9]

34. Huang explained the significance of DPU technology:

> ***[W]hen you take a large scale problem that spans the whole datacenter – it doesn't fit in a single computer – and you accelerate the computation by several orders of magnitude . . . then the network becomes the problem***, and it needs to be very fast. And so that's the reason our relationship with Mellanox goes back a decade and we've been working with them for quite a long time. The networking problem is much, much more complex than just having faster and faster networking. ***And the reason for that is because of the amount of data that you are transmitting, synchronizing, collecting, and reducing across this distributed data center-scale computer and the computation on the fabric itself is complicated.*** . . . ***Putting intelligence in the network – and processing in the network – is vitally important to performance.***[10]

35. Huang also described the importance of offloading to a DPU, just as disclosed in the Xockets Patents:

> A lot of datacenters today have every single packet that is transmitted secured because you want to reduce the attack surface of the datacenter until it's basically every single transaction. There's no way you going to do that on the CPU. ***So you have to move the networking stack off. You want to move the security stack off and you want to move the data processing and data movement stack off.*** And this is something that you want to do right at the NIC before it even comes into the computer and at the NIC before it leaves the computer. The onion, celery, and carrots – you know, ***the holy trinity of computing*** soup – is the CPU, the GPU, and the DPU. . . . ***A DPU is going to be programmable, it's going to do all of that processing that you and I have already talked about, and it's going to offload the movement of data into the granular processing of the data as it's being transmitted and keep it from***

[9] https://www.hpcwire.com/2019/03/14/why-nvidia-bought-mellanox-future-datacenters-will-belike-high-performance-computers.

[10] https://www.nextplatform.com/2020/04/27/nvidia-plus-mellanox-talking-datacenter-architecture-with-jensen-huang.

> ***ever bothering the CPUs and GPUs*** and avoid redundant copies of data. ***<u>That's the architecture of the future</u>.*** And that's the reason why we're so excited about Mellanox.[11]

36. Further explaining the importance of Xockets' revolutionary DPU architecture, Huang declared:

> One of the most important things to disaggregate out of the server node and its CPU is the data processing. That is a giant amount of unnecessary CPU cores running unnecessary software in the datacenter. I don't know how much – its maybe 30 percent to 50 percent. . . . ***I really do think that when you offload [to] the data processing on the SmartNIC***, when you're able to disaggregate the converged server, ***when you can put accelerators anywhere in datacenter*** and then can compose and reconfigure that datacenter for this specific workload – ***<u>that's a revolution</u>***.[12]

37. NVIDIA explained that "DPUs are an essential element of modern and secure data centers in which CPUs, GPUs and DPUs are able to combine into a single computing unit that's fully programmable, AI-enabled and can deliver levels of security and compute power not previously possible."[13] Similarly, in an April 2021 press release, NVIDIA stated that "[a] new type of processor, designed to process data center infrastructure software, is needed to offload and accelerate the tremendous compute load of virtualization, networking, storage, security and other cloud-native AI services. The time for BlueField DPU has come."[14] NVIDIA has described the DPU as a "new pillar" that is "designed to offload, accelerate, and isolate infrastructure workloads

[11] https://www.nextplatform.com/2020/04/27/nvidia-plus-mellanox-talking-datacenter-architecture-with-jensen-huang.

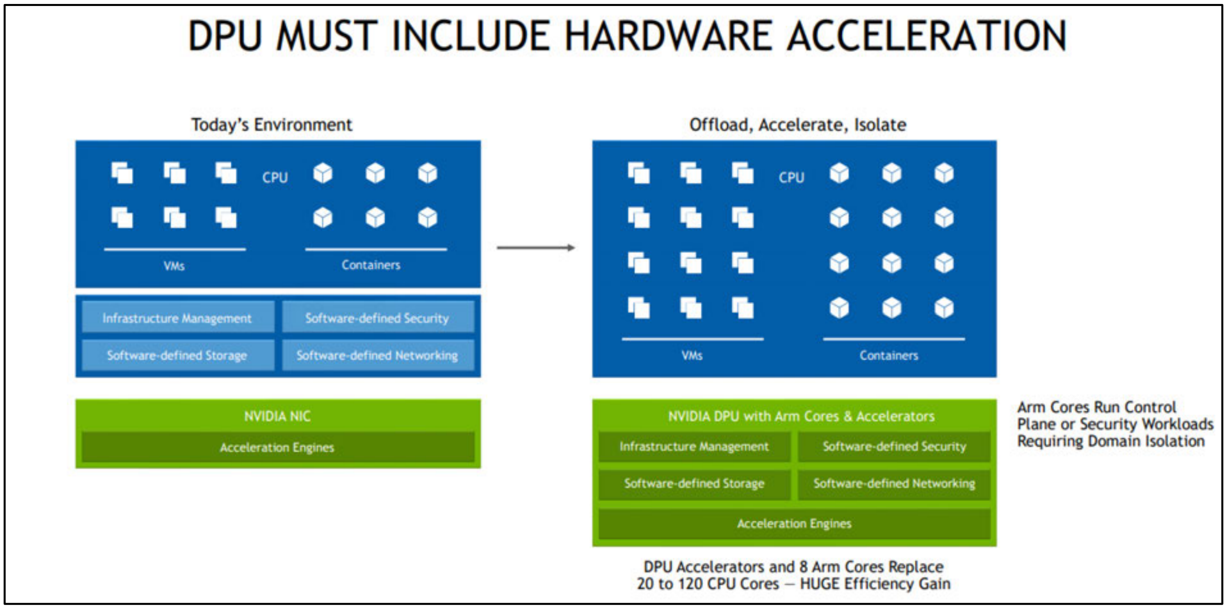
[12] https://www.nextplatform.com/2020/04/27/nvidia-plus-mellanox-talking-datacenter-architecture-with-jensen-huang.

[13] https://nvidianews.nvidia.com/news/nvidia-introduces-new-family-of-bluefield-dpus-to-bring-breakthrough-networking-storage-and-security-performance-to-every-data-center.

[14] https://nvidianews.nvidia.com/news/nvidia-extends-data-center-infrastructure-processing-roadmap-with-bluefield-3.

and bring efficiency and security to software defined workloads such as networking security and storage while freeing CPU resources by up to 30%."[15]

38. In fact, NVIDIA illustrates how "[o]ffloading infrastructure tasks to the DPU improves server performance, efficiency, and security"[16] using Xockets' DPU computing architecture:



DPU MUST INCLUDE HARDWARE ACCELERATION

Today's Environment

CPU

VMs

Containers

Infrastructure Management

Software-defined Security

Software-defined Storage

Software-defined Networking

NVIDIA NIC

Acceleration Engines

Offload, Accelerate, Isolate

CPU

VMs

Containers

NVIDIA DPU with Arm Cores & Accelerators

Infrastructure Management

Software-defined Security

Software-defined Storage

Software-defined Networking

Acceleration Engines

Arm Cores Run Control
Plane or Security Workloads
Requiring Domain Isolation

DPU Accelerators and 8 Arm Cores Replace
20 to 120 CPU Cores – HUGE Efficiency Gain

[17]

39. NVIDIA calls this a "fundamental new architecture."[18] NVIDIA's CEO Huang refers to the DPU paradigm as a "fundamental transition" necessitated by the fact that "CPU

[15] NVIDIA DOCA Software Framework, https://www.youtube.com/watch?v=htR19rdBicA.

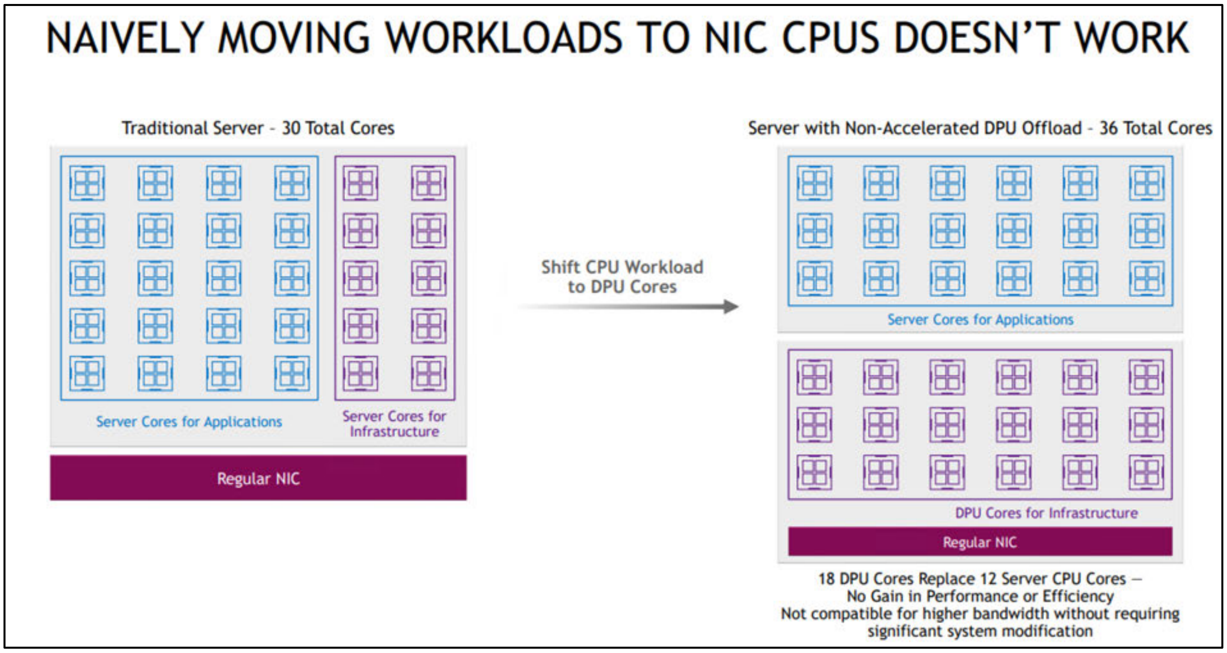
[16] https://developer.nvidia.com/blog/offloading-and-isolating-data-center-workloads-with-bluefield-dpu.

[17] https://hc33.hotchips.org/assets/program/conference/day1/HC2021.NVIDIA.IdanBurstein.v08.norecording.pdf.

[18] https://nvidianews.nvidia.com/news/nvidia-extends-data-center-infrastructure-processing-roadmap-with-bluefield-3 ("'Modern hyperscale clouds are driving a fundamental new architecture for data centers,' said Jensen Huang, founder and CEO of NVIDIA.").

scaling [Moore's law] has ended. We need a new computing approach and accelerated computing is the path forward. . . . This way of doing computation is a reinvention from the ground up."[19]

40. Further, NVIDIA illustrates that Xockets' DPU computing architecture is a breakthrough innovation, admitting that "naively moving workloads to NIC CPUs doesn't work"[20]:



NAIVELY MOVING WORKLOADS TO NIC CPUS DOESN'T WORK

Traditional Server - 30 Total Cores

Server Cores for Applications

Server Cores for Infrastructure

Regular NIC

Shift CPU Workload to DPU Cores

Server with Non-Accelerated DPU Offload - 36 Total Cores

Server Cores for Applications

DPU Cores for Infrastructure

Regular NIC

18 DPU Cores Replace 12 Server CPU Cores –
No Gain in Performance or Efficiency
Not compatible for higher bandwidth without requiring significant system modification

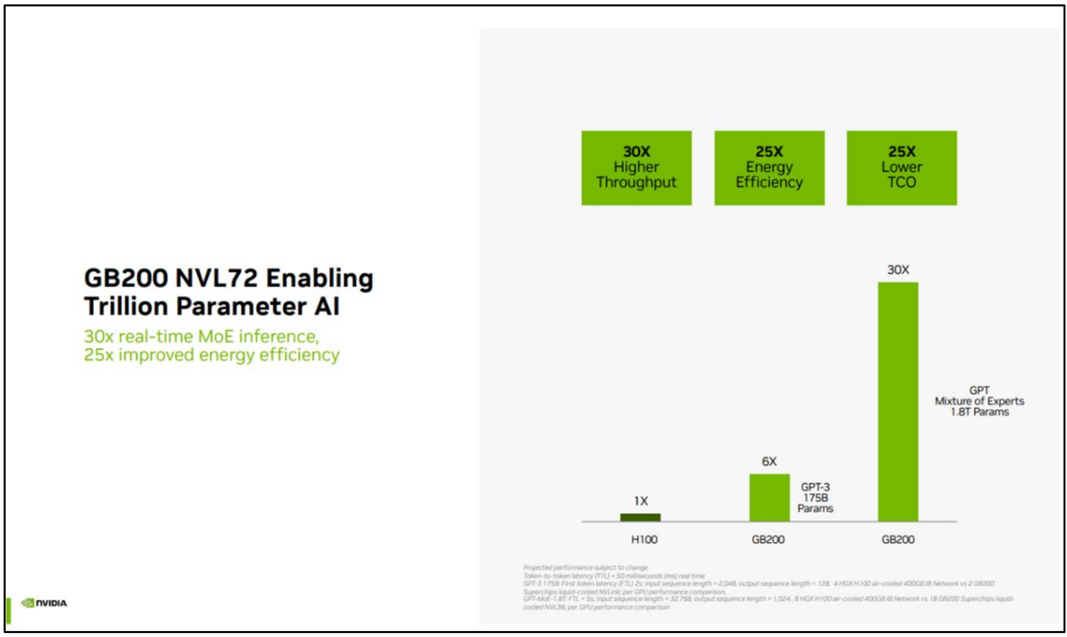
41. While in the media NVIDIA and Huang claimed this technology as NVIDIA's own, Xockets invented, developed, patented, and presented to the industry this technology years earlier.

42. After first learning of NVIDIA's infringement and its collaborations with Microsoft, Dr. Dalal, personally provided NVIDIA with notice of the Xockets Patents on February 10, 2022. NVIDIA did not cease its infringing conduct, or even seek to negotiate for rights to the

[19] https://www.nvidia.com/en-us/events/computex/?nvid=nv-int-cwmfg-130532#cid=cmptx23e_nv-int-cwmfg_en-us; *see also* NVIDIA Keynote at COMPUTEX 2023, https://www.youtube.com/watch?v=i-wpzS9ZsCs&t=875s (14:35–15:47).

[20] https://hc33.hotchips.org/assets/program/conference/day1/HC2021.NVIDIA.IdanBurstein.v08.norecording.pdf.

Xockets Patents. Instead, NVIDIA doubled-down on its infringing conduct by ceasing further contact with Dr. Dalal and exponentially expanding its infringement with the release of its NVLink Switch DPUs; to accelerate the training of large models for AI, including in extraordinary ways with its Blackwell GPU-enabled server computer systems to be released this Fall 2024.[21]



GB200 NVL72 Enabling Trillion Parameter AI

30x real-time MoE inference, 25x improved energy efficiency

Projected performance subject to change
Token-to-token latency (TTL) = 50 milliseconds (ms) real time
GPT-3 175B: First token latency (FTL) 2s; input sequence length = 2,048, output sequence length = 128, 4 HGX H100 air-cooled 400GB IB Network vs 2 GB200 Superchips liquid-cooled NVLink; per GPU performance comparison,
GPT-MoE-1.8T: FTL = 5s; input sequence length = 32,768, output sequence length = 1,024 , 8 HGX H100 air-cooled 400GB IB Network vs 18 GB200 Superchips liquid-cooled NVL36; per GPU performance comparison

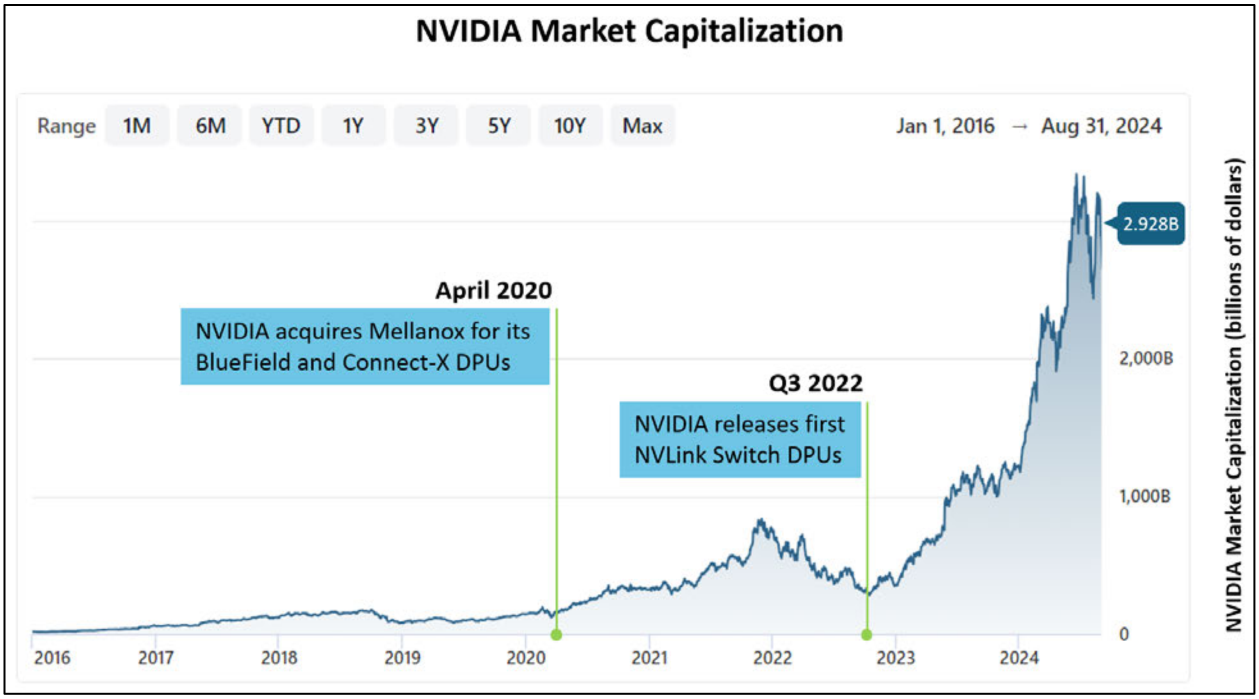
[22]

43. Notably, since the launch of its systems that use Xockets' DPU technology, NVIDIA's market capitalization has surged from approximately $180 billion in April 2020 (when NVIDIA acquired Mellanox and released its BlueField and ConnectX DPUs) to approximately $3 ***trillion*** as of the end of August 2024 (following the release of its NVLink DPUs)[23]:

[21] https://nvidianews.nvidia.com/news/nvidia-announces-dgx-h100-systems-worlds-most-advanced-enterprise-ai-infrastructure ("NVIDIA DGX H100 systems, DGX PODs and DGX SuperPODs will be available from NVIDIA's global partners starting in the third quarter [of 2022]."); https://developer.nvidia.com/blog/?p=53977.

[22] https://www.servethehome.com/nvidia-blackwell-platform-at-hot-chips-2024/nvidia-blackwell-hot-chips-2024_page_26.

[23] https://stockanalysis.com/stocks/nvda/market-cap (Aug. 2024).



44. Huang and NVIDIA's own admissions, examples of which are above, demonstrate that NVIDIA vaulted to become one of the world's most valuable corporations by market capitalization in meaningful part through its widespread implementation and deployment of Xockets' patented DPU technology, described in and protected by Xockets' New Cloud Processor Patents and New Cloud Fabric Patents.

THE PARTIES

I. XOCKETS

45. Xockets is a Texas corporation, with its principal place of business located in the Temple Office Park at 2027 South 61st Street, Suite 107, Temple, Texas 76504. Temple's thriving health and science industry[24] is an environment that Xockets believes will facilitate development of applications of DPU technology and ML/AI to improve people's health and lives.

[24] *See* https://templeedc.com/why-temple-tx-offers-a-healthy-ecosystem-for-healthcare-businesses.

46. Xockets is the developer and owner of foundational technology used to make today's AI breakthroughs possible.

47. Xockets was founded in 2012 by Dr. Parin Dalal.

48. Funding for Xockets was provided by notable investors, including the current CTO of Intel, Dr. Greg Lavender, who was a long-time Xockets Board member, Robert Cote, one of the nation's leading IP investors and lawyers, and also a Xockets Board member, and Jerry Yang, cofounder of Yahoo.

49. Xockets specializes in the development of infrastructure products for distributed computing within the technology sector, including the integration of hardware and software acceleration into appliances for distributed computing, including AI. Its developments are designed to enhance the performance of open source frameworks, reduce power consumption, and lower capital costs, all while being compatible with commodity scale-out data centers.

50. Xockets is the assignee and owns all right, title, and interest to the New Cloud Processor Patents (i.e., the '209 Patent – DPU Computing Architecture, Security; '924 Patent – DPU Network Overlay, Security; and '350 Patent – DPU Stream Processing) and the New Cloud Fabric Patents (i.e., the '297 Patent – DPU Cloud Network Fabric; '161 Patent – DPU Cloud Network Fabric; '092 Patent – DPU In-Network Computing; and '640 Patent – DPU In-Network Computing).

II. NVIDIA

51. Defendant NVIDIA is a Delaware corporation. NVIDIA is registered with the State of Texas and may be served with process through its registered agent, Corporation Service Company d/b/a CSC-Lawyers Incorporating Service Company, 211 E. 7th Street, Suite 620, Austin, Texas 78701. NVIDIA maintains a facility in Austin at 11001 Lakeline Boulevard, Suite #100 Building 2, Austin, Texas 78717.

III. MICROSOFT

52. Defendant Microsoft Corporation is a Washington corporation. Microsoft is registered with the State of Texas and may be served with process through its registered agent, Corporation Service Company d/b/a CSC-Lawyers Incorporating Service Company, 211 E. 7th Street, Suite 620, Austin, Texas 78701. Microsoft maintains a facility in Austin at 10900 Stonelake Blvd, Suite 225, Austin, Texas 78759.

IV. RPX

53. Defendant RPX Corporation is a Delaware corporation. RPX is registered with the State of Texas and may be served with process through its registered agent, Incorporating Services, Ltd., 3610-2 North Josey, Suite 223, Carrolton, Texas 75007.

JURISDICTION AND VENUE

54. This Court has subject matter jurisdiction over federal antitrust claims pursuant to 15 U.S.C. §§ 15 and 26 and 28 U.S.C. § 1331.

I. NVIDIA

55. NVIDIA is subject to this Court's personal jurisdiction consistent with the principles of due process and/or the Texas Long Arm Statute. Personal jurisdiction exists generally over NVIDIA because NVIDIA has sufficient minimum contacts and/or has engaged in continuous and systematic activities in the forum as a result of business conducted within Texas, including in the Western District of Texas. For example, on information and belief, NVIDIA has committed, and continues to commit, violations of federal antitrust laws in the State of Texas and this District; NVIDIA purposefully availed itself of the privileges of conducting business in the State of Texas and this District; and NVIDIA regularly conducts and solicits business within the State of Texas and this District.

56. Furthermore, personal jurisdiction over NVIDIA in this action comports with due process. For example, NVIDIA has conducted and regularly conducts business within this District; NVIDIA has purposefully availed itself of the privileges of conducting business in this District; and NVIDIA has sought protection and benefit from the laws of the State of Texas. Having purposefully availed itself of the privilege of conducting business within this District, NVIDIA should reasonably and fairly anticipate being brought into court here.

57. NVIDIA has repeatedly acknowledged this Court has personal jurisdiction over it. *See, e.g.*, *Vantage Micro LLC v. NVIDIA Corporation*, Case No. 6:19-cv-00582-RP, Dkt. 22 (W.D. Tex., Jan. 4, 2020) (admitting to personal jurisdiction); *Ocean Semiconductor LLC v. NVIDIA Corporation*, Case No. 6:20-cv-01211-ADA, Dkt. 14 (W.D. Tex., Mar. 12, 2021) (same). Further, NVIDIA has admitted "it is subject to this Court's general personal jurisdiction." *Id.*

58. Venue is proper in the Western District of Texas pursuant to 28 U.S.C. §§ 1391(b)-(d) and/or 15 U.S.C. § 22, including but not limited to because NVIDIA has a regular and established place of business in this District through which it transacts business and where it may be found. Further, as detailed herein, NVIDIA is subject to this Court's personal jurisdiction with respect to the violations described herein.

II. MICROSOFT

59. Microsoft is subject to this Court's personal jurisdiction consistent with the principles of due process and/or the Texas Long Arm Statute. Personal jurisdiction exists generally over Microsoft because Microsoft has sufficient minimum contacts and/or has engaged in continuous and systematic activities in the forum as a result of business conducted within Texas, including in the Western District of Texas. For example, on information and belief, Microsoft has committed, and continues to commit, violations of federal antitrust laws in the State of Texas and this District; Microsoft purposefully availed itself of the privileges of conducting business in the

State of Texas and this District; and Microsoft regularly conducts and solicits business within the State of Texas and this District.

60. In addition, on information and belief, Microsoft has not disputed personal jurisdiction in cases filed against it in this District. *See, e.g.*, *Panther Innovations v. Microsoft Corp.*, No. 6:20-cv-01071, Dkt. No. 14; *Exafer Ltd v. Microsoft Corp.*, No. 1:20-cv-00131, Dkt. No 15; *WSOU Investments, LLC v. Microsoft Corp.*, No. 6:20-cv-00464, Dkt. No. 20; *Zeroclick, LLC v. Microsoft Corp.*, No. 1:20-cv-00272, Dkt. No. 14.

61. Furthermore, personal jurisdiction over Microsoft in this action comports with due process. For example, on information and belief, Microsoft has conducted and regularly conducts business within this District; Microsoft has purposefully availed itself of the privileges of conducting business in this District; and Microsoft has sought protection and benefit from the laws of the State of Texas. Having purposefully availed itself of the privilege of conducting business within this District, Microsoft should reasonably and fairly anticipate being brought into court here.

62. Venue is proper in the Western District of Texas pursuant to 28 U.S.C. §§ 1391(b)-(d) and/or 15 U.S.C. § 22, including but not limited to because Microsoft has a regular and established place of business in this District through which it transacts business and where it may be found. Further, as detailed herein, Microsoft is subject to this Court's personal jurisdiction with respect to the violations described herein.

III. RPX

63. RPX is subject to this Court's personal jurisdiction consistent with the principles of due process and/or the Texas Long Arm Statute. Personal jurisdiction exists generally over RPX because RPX has sufficient minimum contacts and/or has engaged in continuous and systematic activities in the forum as a result of business conducted within Texas, including in the Western

District of Texas. For example, on information and belief, RPX has committed, and continues to commit, violations of federal antitrust laws in the State of Texas and this District; RPX purposefully availed itself of the privileges of conducting business in the State of Texas and this District; and RPX regularly conducts and solicits business within the State of Texas and this District.

64. Furthermore, personal jurisdiction over RPX in this action comports with due process. For example, on information and belief, RPX has conducted and regularly conducts business within this District; RPX has purposefully availed itself of the privileges of conducting business in this District; and RPX has sought protection and benefit from the laws of the State of Texas. Having purposefully availed itself of the privilege of conducting business within this District, RPX should reasonably and fairly anticipate being brought into court here.

65. Venue is proper in the Western District of Texas pursuant to 28 U.S.C. §§ 1391(b)-(d) and/or 15 U.S.C. § 22, including but not limited to because RPX transacts business in this District. Further, as detailed herein, RPX is subject to this Court's personal jurisdiction with respect to the violations described herein.

FACTUAL ALLEGATIONS

I. XOCKETS PATENTS

66. Xockets' New Cloud Processor Patents (including the '209, '924, and '350 Patents) and its New Cloud Fabric Patents (including the '297, '161, '092, and '640 Patents) are exemplary Xockets patents that are targets of the illegal buyers' cartel discussed herein.

A. THE NEW CLOUD PROCESSOR PATENTS

67. Xockets invented the DPU and its virtual switch computing architecture years before the industry to enable accelerated computing and AI in cloud data centers, providing the versatility needed in offloading, accelerating, and isolating from cloud server processors (e.g.,

CPUs, GPUs, and hybrids of these server processors) the data-intensive computing tasks required to make distributed computing in data centers possible, including for training large models for AI.

68. Xockets' DPU computing architecture is protected by the New Cloud Processor Patents, including the '209, '924, and '350 Patents, discussed below.

(1) The '209 Patent – DPU Computing Architecture, Security

69. U.S. Patent No. 11,080,209 ("the '209 Patent") is entitled "Server Systems and Methods for Decrypting Data Packets With Computation Modules Insertable Into Servers That Operate Independent of Server Processors." The '209 Patent duly and legally issued on August 3, 2021, from U.S. Patent Application No. 15/396,334, filed on December 30, 2016.

70. The '209 Patent is a continuation of U.S. Patent Application No. 13/900,346, filed on May 22, 2013, and claims priority from U.S. Provisional Application No. 61/650,373, filed on May 22, 2012. The '209 Patent is entitled to the benefit of these earlier filed applications.

71. Xockets is the current owner of all rights, title, and interest in and to the '209 Patent, including the right to sue for past damages.

72. A true and correct copy of the '209 Patent is attached hereto as **Exhibit 1** and is incorporated by reference herein.

73. The '209 Patent relates to server systems in cloud data centers utilizing a novel computing architecture in a new cloud processor, or DPU, for offloading, accelerating, and isolating data-intensive computing operations from server processors (CPUs, GPUs, and hybrids of these server processors), including for cloud infrastructure services and big data analytics applications such as those used in training large language models for AI. The server system can include a plurality of servers interconnected by a network. Each server includes a server processor that is configured to execute an operating system for the server. Each server further includes a computation module, or DPU, that is separate from the server processor and is coupled to the server

processor by a bus. The computation module, or DPU, includes a virtual switch computing architecture for identifying and classifying packet flows, also known as sessions, and connecting together identified programmable logic pipelines of hardware accelerators or computation elements, comprising offload processors or offload processing circuits. The computation module, or DPU, uses packet data to define in the virtual switch the programmable logic pipelines to be formed for computational operations, or what is referred to as data-centric computing. This data-centric computing approach supports cloud computing at the speed of the network, or line rate. The programmable hardware acceleration formed using the virtual switch is for performing computing operations on packet data independent of server processors in the cloud data center.

74. The invention of the ’209 Patent solves a technological problem with prior art server systems and methods for performing computationally intensive workloads such as processing of packets for high-volume applications. For example, the ’209 Patent explains that “[p]acket handling and security applications can require a significant amount of scarce computational resources in enterprise server or cloud based data systems.” ’209 Patent, 1:27–29.

75. Conventional approaches of throwing more hardware at the problem under control of server processors, for processing data-intensive workloads and sources in cloud data centers, were too expensive and did not address the root cause of the problem, which was that the server processors in cloud servers were constantly interrupted and bottlenecked with data-intensive infrastructure services such as for “packet handling and transport services.” *Id.*, 1:32–38. These conventional architectures were thus ill-equipped to handle such high-volume applications: “Even idling, x86 processors use a significant amount of power, and near continuous operation for high bandwidth packet analysis functionality make the processor energy costs one of the dominant price factors.” *Id.*, 1:39–44. “In addition, issues with the high cost of context switching, the limited

parallelism, and the security implications associated with running encryption/decryption modules on x86 processors have reduced the effectiveness of enterprise or cloud data security." *Id.*, 1:44–48.

76. The '209 Patent improved upon the systems and methods in the prior art by introducing a new cloud processor located at the boundary of the network leading to each server processor, providing bump-in-the-wire hardware acceleration for data-intensive computing operations independent of server processors using the virtual switch computing architecture. As described in the '209 Patent, this Xockets virtual switch is programmed to form programmable logic pipelines of hardware acceleration and has the versatility needed for cloud adoption. Such cloud processors or DPUs are referred to as computation modules or offload processing modules in the '209 Patent and in one embodiment are referred to as Xocket In-line Memory Modules ("XIMMs"). *See id.*, 2:13–20. "Using one or more XIMMs it is possible to execute lightweight packet handling tasks without intervention from a main server processor." *Id.*, 2:20–22.

77. The invention of the '209 Patent "can have high efficiency context switching, high parallelism, and can solve security problems associated with running encryption/decryption modules on x86 processors. Such systems as a whole are able to handle high network bandwidth traffic at a lower latency and at a very low power when compared to traditional high power 'brawny' server cores." *Id.*, 2:22–29. The invention of the '209 Patent can thus provide software-defined hardware acceleration in processing data-intensive workloads of cloud infrastructure services with lower power costs and high reliability. *See id.*, 2:29–33. A variety of cloud infrastructure services can thus be offloaded to the computation module and accelerated, and run independent of server processors at the line rate of the network, providing the versatility needed to offload and accelerate various cloud applications and infrastructure services "including but not

limited to virtual private network (VPN) tunneling and signature detection and packet filtering as an intrusion prevention system (IPS)" such as used in cloud VPN communications, providing levels of compute and security in cloud data centers that were not previously possible. *See id.*, 2:46–50.

78. For example, Claim 18 of the '209 Patent is directed to:

> 18. A server system, comprising:
>
> a plurality of servers interconnected by a network, each server including
>
> > a server processor configured to execute an operating system for the server,
> >
> > at least one computation module, separate from the server processor and coupled to the server processor by at least one bus, the at least one computation module including
> >
> > > first processing circuits mounted on the computation module and configured to
> > >
> > > > execute header detection on packets received by the server,
> > > >
> > > > classifying received packets by a session identifier, and
> > > >
> > > > operate as a virtual switch to provide packets to circuits on the at least one computation module, and
> > >
> > > at least decryption circuits implemented on programmable logic devices and configured to decrypt received packets; wherein
>
> the computation modules execute header detection, classifying of packets, virtual switching of packets, and decryption of packets independent of the server processor of their respective server.

79. For example, Claim 20 of the '209 Patent is directed to:

> 20. The server system of claim 18, wherein the at least decryption circuits decrypt the received packets according to a virtual private network (vpn) encryption/decryption protocol.

80. NVIDIA is not licensed to the '209 Patent.

81. Microsoft is not licensed to the ’209 Patent.

(2) The ’924 Patent – DPU Network Overlay, Security

82. U.S. Patent No. 10,649,924 (“the ’924 Patent”) is entitled “Network Overlay Systems and Methods Using Offload Processors.” The ’924 Patent duly and legally issued on May 12, 2020, from U.S. Patent Application No. 15/396,323, filed on December 30, 2016.

83. The ’924 Patent is a continuation of U.S. Patent Application No. 13/921,059, filed on June 18, 2013, and claims priority from U.S. Provisional Application Nos. 61/753,901; 61/753,906; 61/753,892, 61/753,899; 61/753,903; 61/753,895; 61/753,910; 61/753,904; and 61/753,907, all filed on January 17, 2013. The ’924 Patent is entitled to the benefit of these earlier filed applications.

84. Xockets is the current owner of all rights, title, and interest in and to the ’924 Patent, including the right to sue for past damages.

85. A true and correct copy of the ’924 Patent is attached hereto as **Exhibit 2** and is incorporated by reference herein.

86. The ’924 Patent relates to systems, hardware, and methods in cloud data centers utilizing a virtual switch computing architecture in a new cloud processor, or DPU, to offload, accelerate, and isolate data-intensive computing operations from server processors (CPUs, GPUs, and hybrids of these server processors) to provide network overlay infrastructure services for improved cloud security, including by implementing network overlays for cloud VPN communications. In particular, the ’924 Patent relates to network overlay services that are provided by the new cloud processor, or DPU, also called offload processor modules, that receives data packets and routes them to programmable logic pipelines of hardware accelerators comprising offload processors or offload processing circuits for packet encapsulation, decapsulation, modification, or data handling, such as in cloud VPN communications. The offload processor

modules are mounted to a system bus of a host server, that further includes a host processor connected to the system bus. Offload processor modules include offload processors or offload processing circuits that function as hardware accelerators and are configured to encapsulate network packet data for transport on a logical network or decapsulate the network packet data received from the logical network. The offload processing circuits encapsulate or decapsulate network packet data independent of any host processor and provide programmable hardware acceleration to packet encapsulation and decapsulation functions. This is critical to enabling cloud data centers to offload the provisioning for each customer a secure virtual network that is seemingly a different network to the customer, but is actually running on the same physical network.

87. The '924 Patent discloses and claims improved systems and methods for processing packets in cloud data centers using network overlay services. "Modern computing systems can support a variety of intercommunication protocols. In certain instances, computers can connect with each other using one network protocol, while appearing to outside users to use another network protocol. Commonly termed an 'overlay' network, such computer networks are effectively built on top of another computer network, with nodes in the overlay network being connected by virtual or logical links to the underlying network." '924 Patent, 1:27–34.

88. While "[o]verlay networks are particularly useful for environments where different physical network servers, processors, and storage units are used, and network addresses to such devices may commonly change," "overlay networks do require additional computational processing power to run." *Id.*, 1:47–55. Therefore, "efficient network translation mechanisms are necessary, particularly when large numbers of network transactions occur." *Id.*, 1:55–57.

89. To solve this issue, the '924 Patent provides improved systems, hardware, and methods that allow for "high speed and/or energy efficient processing of packet data that does not necessarily require access to computing resources of a host processor of a server, server rack system, or blade server." *Id.*, 1:61–65. Instead, packets can be directed to and processed by offload processor modules, or DPUs, which can operate on the packet data independent of any host processors. Thus, using the invention of the '924 Patent, "server loads can be broken up across the offload processing cores and the host processing cores." *Id.*, 5:19–21.

90. As a result, the invention of the '924 Patent "can provide improved computational performance as compared to traditional computing systems," in providing Cloud VPN services, which "are often ill-equipped to handle such high volume applications." *Id.*, 8:37–41.

91. For example, Claim 9 of the '924 Patent is directed to:

> 9. A method for providing network overlay services, comprising the steps of:
>
> receiving network packet data from a data source in an offload processor module that is mounted to a system bus of a host server, the host server further including
>
> > at least one host processor connected to the system bus, and
> >
> > a network interface device;
>
> encapsulating the network packet data to create encapsulated network packets for transport on a logical network or decapsulating the network packet data to create decapsulated network packets for delivery to a network location, the encapsulating and decapsulating being executed by processing circuits mounted on the offload processor module and being executed independent of any host processor; and
>
> transporting the encapsulated network packets or the decapsulated network packets out of the offload processor module; wherein
>
> > the logical network is overlaid on a physical network.

92. NVIDIA is not licensed to the '924 Patent.

93. Microsoft is not licensed to the ’924 Patent.

(3) The ’350 Patent – DPU Stream Processing

94. U.S. Patent No. 11,082,350 (“the ’350 Patent”) is entitled “Network Server Systems, Architectures, Components and Related Methods” for expanding the versatility of a DPU for cloud offload by adding general-purpose processors (e.g., ARM cores) with a modified architecture described in the ’350 Patent that ensures that they can function as general-purpose hardware accelerators and process data-intensive workloads at the speed of the network, or line rate. The ’350 Patent duly and legally issued on August 3, 2021, from U.S. Patent Application No. 16/129,762, filed on September 12, 2018.

95. The ’350 Patent is a continuation-in-part of U.S. Patent Application No. 15/396,318, filed on December 30, 2016, which is a continuation of U.S. Patent Application No. 13/900,318, filed on May 22, 2013, and U.S. Patent Application No. 15/283,287 (“the ’287 Application”), filed on September 30, 2016. Further, the ’287 Application is a continuation of International Patent Application Nos. PCT/US2015/023730 and PCT/US2015/023746, filed on March 31, 2015. In addition, the ’350 Patent claims priority from U.S. Provisional Application Nos. 62/557,659; 62/557,661; 62/557,666; 62/557,670; 62/557,671; 62/557,675; 62/557,679; 62/557,687, all filed on September 12, 2017; U.S. Provisional Application No. 61/976,471, filed on April 7, 2014; U.S. Provisional Application Nos. 61/973,207 and 61/973,205, filed on March 31, 2014; U.S. Provisional Application Nos. 61/753,892; 61/753,895; 61/753,901; 61/753,903; 61/753,904; 61/753,906; 61/753,910; 61/753,907; and 61/753,899, all filed on January 17, 2013; and U.S. Provisional Application No. 61/650,373, filed on May 22, 2012. The ’350 Patent is entitled to the benefit of these earlier filed applications.

96. Xockets is the current owner of all rights, title, and interest in and to the ’350 Patent, including the right to sue for past damages.

97. A true and correct copy of the ’350 Patent is attached hereto as **Exhibit 3** and is incorporated by reference herein.

98. The ’350 Patent relates to systems in cloud data centers utilizing a novel computing architecture in a new cloud processor, or DPU, also called hardware acceleration modules, to offload, accelerate, and isolate data-intensive computing operations from server processors (CPUs, GPUs, and hybrids of these server processors). The system includes a server with a host processor and at least one hardware acceleration (hwa) module, or DPU, physically separate from the host processor. Hardware accelerators or computing elements are formed in programmable logic pipelines on the hardware acceleration module that include offload processors or offload processing circuits configured to execute a plurality of processes, first memory circuits, and second memory circuits. The computing elements further include a hardware scheduler circuit for streaming network packet flows to the processing circuits (e.g., general-purpose processors such as ARM cores) using their associated memory circuits so that they function like hardware accelerators. The hardware acceleration module includes a data transfer fabric configured to enable data transfers between the processing circuits and the first and second memory circuits; wherein the computing elements are configured to transfer data to, or receive data from, any of: the processing circuits, the first memory circuits, the second memory circuits, or other computing elements coupled to the data transfer fabric. The addition of the hardware scheduler enables stream processing in the hardware acceleration modules, providing run-to-completion computational processing of packet flows in general-purpose processors that now function like hardware accelerators at the speed of the network, or line rate.

99. The invention of the ’350 Patent solves a technological problem with prior art data processing systems, including systems and methods for processing large data sets. For example,

the ’350 Patent explains that “[c]onventional data intensive computing platforms for handling large volumes of unstructured data can use a parallel computing approach combining multiple processors and disks in large commodity computing clusters connected with high-speed communications switches and networks.” ’350 Patent, 15:3–7. An exemplary programming model for processing large data sets is known as “map, reduce.” *Id.*, 15:14–17. However, in such conventional systems, “data spills to disk are almost unavailable. This slows performance and such spilled data needs to be read back into server memory to continue processing.” *Id.*, 15:34–37.

100. Accordingly, the ’350 Patent recognized that “[i]t would be desirable to arrive at some way of increasing the performance of [] systems for processing unstructured data that do not suffer from the drawbacks of conventional approaches.” *Id.*, 15:43–46. To that end, the ’350 Patent discloses and claims improved systems and methods, including those “that can perform data processing, including ‘big’ data processing, by accelerating processing tasks with networked hardware accelerator (hwa) modules included in server systems.” *Id.*, 15:47–50. The improved systems “can provide map, reduce type processing, without data skew and/or spills to disk that can occur in conventional architectures.” *Id.*, 18:8–11. As an example, in the case of large language model training for AI requiring “machine learning applications to run across multiple computing elements on multiple networked servers,” the stream processing invention enables in-network or in-flight computing operations such as reduction/combining of training results. *Id.*, 9:12–14.

101. For example, Claim 1 of the ’350 Patent is directed to:

> 1. A device, comprising:
>
> a server that includes a host processor and at least one hardware acceleration (hwa) module physically separate from the host processor and having
>
> > a network interface configured to virtualize functions by redirecting network packets to different addresses within the hwa,

> > at least one computing element formed thereon, the at least one computing element including
> >
> > > processing circuits configured to execute a plurality of processes including at least one virtualized function,
> > >
> > > a scheduler circuit configured to allocate a priority to a processing of packets of one flow over those of another flow by the processing circuits,
> > >
> > > first memory circuits,
> > >
> > > second memory circuits, and
> >
> > a data transfer fabric configured to enable data transfers between the processing circuits and the first and second memory circuits; wherein
>
> the at least one computing element is configured to transfer data to, or receive data from, any of: the processing circuits, the first memory circuits, the second memory circuits, or other computing elements coupled to the data transfer fabric.

102. NVIDIA is not licensed to the ’350 Patent.

103. Microsoft is not licensed to the ’350 Patent.

B. THE NEW CLOUD FABRIC PATENTS

104. Xockets invented a DPU switching architecture for connecting together its DPUs in a novel way to form a “New Cloud Fabric” in data centers—one that can bypass the existing cloud network and its limitations to enable accelerated computing and AI in data centers, and turn every data center into an AI factory.

105. Xockets’ DPU switching architecture for forming a new cloud fabric is protected by the New Cloud Fabric Patents, including the ’297 and ’161 Patents, which claim this New Cloud Fabric, and the ’092 and ’640 Patents, which claim offloading from server processors, accelerating, and isolating the processing of data-intensive workloads in this DPU fabric.

106. Xockets’ New Cloud Fabric enables brokering of high-speed collective communication of data between server processors in a cloud data center and in-network computing

to sort, organize, and reduce the data-intensive workloads involved in training AI models across GPU servers. This enables the higher speeds and power efficiency needed to make AI production affordable and widely available.

(1) The ’297 Patent – DPU Cloud Network Fabric

107. U.S. Patent No. 10,223,297 (“the ’297 Patent”) is entitled “Offloading of Computation for Servers Using Switching Plane Formed by Modules Inserted Within Such Servers” for forming a new cloud fabric that can operate independent of server processors. The ’297 Patent duly and legally issued on March 5, 2019, from U.S. Patent Application No. 15/396,328, filed on December 30, 2016.

108. The ’297 Patent is a continuation of U.S. Patent Application No. 13/900,222, filed on May 22, 2013, and claims priority from U.S. Provisional Application No. 61/650,373, filed on May 22, 2012. The ’297 Patent is entitled to the benefit of these earlier filed applications.

109. Xockets is the current owner of all rights, title, and interest in and to the ’297 Patent, including the right to sue for past damages.

110. A true and correct copy of the ’297 Patent is attached hereto as **Exhibit 4** and is incorporated by reference herein.

111. The ’297 Patent relates to server systems in cloud data centers, and more particularly to computation modules, or DPUs, also called offload processor modules, in such systems that are connected to form a new switching plane or new cloud fabric that can operate independent of server processors. The system includes a plurality of first server modules interconnected to one another via a communication network, wherein each first server module includes a first switch for forming a first switching plane, at least one main processor, and at least one computation module, or DPU, coupled to the main processor by a bus. Each computation module, or DPU, includes a second switch and a plurality of computation elements that function

as hardware accelerators, comprising offload processors or offload processing circuits for performing programmable hardware acceleration in the network. The second switches of the first server modules form a second switching plane or cloud fabric for the ingress and egress of network packets independent of any main processors of the first server modules. Furthermore, the second switches include virtual switches that switch together programmable logic pipelines of hardware accelerators for offloading from server processors data-intensive workloads of a cloud, such as used in collective communication of training data as well as in-network computing operations for reducing/combining that data, which is critical in training large language models for AI.

112. The ’297 Patent solves a technological problem with server systems. For example, the ’297 Patent explains that “[n]etworked applications often run on dedicated servers that support an associated ‘state’ context or session-defined application. Servers can run multiple applications, each associated with a specific state running on the server.” ’297 Patent, 1:23–26. “Unfortunately, servers can be limited by computational and memory storage costs associated with switching between applications. When multiple applications are constantly required to be available, the overhead associated with storing the session state of each application be result in poor performance due to constant switching between applications.” *Id.*, 1:32–38. “Dividing applications between multiple processor cores” does not solve the problem, “since even advanced processors often only have eight to sixteen cores, while hundreds of application or session states may be required.” *Id.*, 1:38–42.

113. To address this issue, the ’297 Patent discloses and claims improved systems and methods with computation modules, or DPUs, also referred to as offload processor modules, that can run such session-defined applications in part or full. *See id.*, 2:10–18. In one embodiment, “[i]n effect, one can reduce problems associated with session limited servers by using the module

processor (e.g., an ARM architecture processor) of a XIMM to offload part of the functionality of traditional servers." *Id.*, 2:48–51.

114. The '297 Patent further explains a number of improvements can be achieved using the claimed switching architecture in one embodiment of a new cloud fabric or switching plane "to ingress and egress packets within a parallel mid-plane formed from XIMMs." *Id.*, 9:25–29. This new switching plane or cloud fabric enables the acceleration of collective communication and reduction/combining operations critical in training large models for AI: "An additional benefit, among others, with such an architecture is the acceleration of Map-Reduce algorithms by an order of magnitude, making them suitable for business analytics." *Id.*, 9:53–56.

115. For example, Claim 1 of the '297 Patent is directed to:

> 1. A system, comprising:
>
> a plurality of first server modules interconnected to one another via a communication network, each first server module including
>
> > a first switch,
> >
> > at least one main processor, and
> >
> > at least one computation module coupled to the main processor by a bus, each computation module including
> >
> > > a second switch, and
> > >
> > > a plurality of computation elements; wherein
>
> the second switches of the first server modules form a switching plane for the ingress and egress of network packets independent of any main processors of the first server modules, and
>
> each computation module is insertable into a physical connector of the first server module.

116. For example, Claim 7 of the '297 Patent is directed to:

> 7. The system of claim 1, wherein the second switch is a virtual switch comprising computation elements on the computation module.

117. NVIDIA is not licensed to the ’297 Patent.

118. Microsoft is not licensed to the ’297 Patent.

(2) The ’161 Patent – DPU Cloud Network Fabric

119. U.S. Patent No. 9,378,161 (“the ’161 Patent”) is entitled “Full Bandwidth Packet Handling With Server Systems Including Offload Processors” for forming a new switching plane or cloud fabric that can overcome the speed limitations of the existing cloud network and server systems. The ’161 Patent duly and legally issued on June 28, 2016, from U.S. Patent Application No. 13/931,903 (“the ’903 Application”), filed on June 29, 2013.

120. The ’903 Application was a substitute for U.S. Patent Application No. 61/753,892, filed on January 17, 2013. In addition, the ’161 Patent claims priority from U.S. Provisional Application Nos. 61/753,895; 61/753,899; 61/753,901; 61/753,903; 61/753,904; 61/753,906; 61/753,907; and 61/753,910, all filed on January 17, 2013. The ’161 Patent is entitled to the benefit of these earlier filed applications.

121. Xockets is the current owner of all rights, title, and interest in and to the ’161 Patent, including the right to sue for past damages.

122. A true and correct copy of the ’161 Patent is attached hereto as **Exhibit 5** and is incorporated by reference herein.

123. The ’161 Patent relates to improved systems, hardware, and methods for cloud data centers by creating a rack level server system having offload processor modules, or DPUs, connected together in a novel switching architecture to form a new switching plane or cloud fabric. In particular, the ’161 patent relates to rack level or cluster level server systems that include a plurality of servers mountable in a rack and a top of rack (TOR) unit having connections to each of the servers and the existing cloud network. A plurality of offload processor modules are disclosed for offloading data-intensive workloads from server processors of the rack-level server

system. Each offload processor module includes multiple offload processors that function as programmable hardware accelerators and at least one input-output (IO) port. The offload processor modules are connected directly to each other through their respective IO ports and to memory of each server to form a new cloud fabric for cloud offload from server processors, that can bypass the limitations associated with collective communication over the existing cloud network and server systems including conventional Top-Of-Rack switches prone to congestion.

124. The ’161 Patent solves a technological problem with server systems., including systems used in data centers and for data processing. For example, the ’161 Patent explains that “[e]fficient managing of network packet flow and processing is critical for high performance,” and “[s]ubstantial improvements in network service would be made possible by systems that can flexibly process a data flow, recognize or characterize patterns in the data flow, and improve routing and processing decisions for the data.” ’161 Patent, 1:26–35. “Unfortunately, the tree-like server connection topology often used in conventional data centers can be prone to traffic slowdowns and computational bottlenecks.” *Id.*, 1:36–38. “Typically, all the servers in such data centers communication with each other through higher level Ethernet-type switches, such as Top-Of-Rack (TOR) switches.” *Id.*, 1:38–40. However, as the ’161 Patent explains, “[f]low of all the traffic through such TOR switches leads to congestion results in increased network latency, particularly during the periods of high usage.” *Id.*, 1:41–43.

125. Improving upon the prior art, the ’161 Patent discloses and claims systems, hardware, and methods relating to rack server systems having DPUs connected together in a novel switching architecture to form a new switching plane or cloud fabric. For example, “to prevent data transfer bottlenecks through TOR switches, and/or to improve[] system performance,” the ’161 Patent discloses and claims systems in which “direct inter-rack and/or intra-rack

communication can be enabled by offload processor modules included in the servers," bypassing the limitations associated with the existing cloud network and server systems. *Id.*, 4:19–23. "[S]uch data communication via offload processor modules can require less time and/or less processing power as compared to TOR switching via aggregation layer transfers. Accordingly, such data transfers can be executed in a more efficient manner than conventional systems." *Id.*, 4:29–34. In addition, "[a]dvantageously, inter/intra-rack communications via offload server modules can also reduce the need for additional TOR switches and can be included to increase bandwidth and introduce redundancy, particularly since TOR switches may have to be periodically replaced to handle higher network speeds." *Id.*, 4:35–40.

126. For example, Claim 1 of the '161 Patent is directed to:

> 1. A rack server system for a packet processing, comprising:
>
> a plurality of servers mountable in a rack;
>
> a top of rack (TOR) unit having connections to each of the servers;
>
> a plurality of offload processor modules, each offload processor module having at least one input-output (IO) port and multiple offload processors, including at least a first offload processor module connected directly to a second offload processor module through their respective IO ports, the offload processor modules are connected to a memory bus on each of the servers, and are further configured to receive network packets from the server through the memory bus and from the IO port on the offload processing module; and
>
> a memory controller configured to send network packet data directly to at least one offload processor module via the memory bus to which the offload processor module is attached.

127. NVIDIA is not licensed to the '161 Patent.

128. Microsoft is not licensed to the '161 Patent.

(3) The ’092 Patent – DPU In-Network Computing

129. U.S. Patent No. 10,212,092 (“the ’092 Patent”) is entitled “Architectures and Methods for Processing Data in Parallel Using Offload Processing Modules Insertable Into Servers” for in-network computing operations on data-intensive workloads in the new switching plane or new cloud fabric of the ’297 and ’161 Patents. The ’092 Patent duly and legally issued on February 19, 2019, from U.S. Patent Application No. 15/396,330, filed on December 30, 2016.

130. The ’092 Patent is a continuation of U.S. Patent Application No. 13,900,318, filed on May 22, 2013, and claims priority from U.S. Provisional Application No. 61/650,373, filed on May 22, 2012; and U.S. Provision Application Nos. 61/753,892; 61/753,895; 61/753,899; 61/753,901; 61/753,903; 61/753,904; 61/753,906; 61/753,907; and 61/753,910, all filed on January 17, 2013. The ’092 Patent is entitled to the benefit of these earlier filed applications.

131. Xockets is the current owner of all rights, title, and interest in and to the ’092 Patent, including the right to sue for past damages.

132. A true and correct copy of the ’092 Patent is attached hereto as **Exhibit 6** and is incorporated by reference herein.

133. The ’092 Patent relates to a cloud distributed computing architecture for executing at least first and second computing operations in parallel for processing data-intensive workloads of server processors, including CPUs, GPUs, or hybrids of these server processors. The distributed computing architecture includes a plurality of servers, each server having an offload processing module, or DPU, and a virtual switch along with computation elements that function as hardware accelerators that can be formed into programmable logic pipelines by the virtual switch to offload data-intensive workloads from server processors for executing the second computing operations. These second computing operations can include, for example, collective communication of training results in training large language models for AI as well as in-network reduction/combining

of the training results in the new switching plane or cloud fabric that is formed with the offload processing modules, or DPUs. The reduction/combining of training results structures the data for use by the GPUs in further training of large language models. These second computing operations are performed on first processed data in the claimed invention that can include, for example, the training results that are generated by first computing operations on the GPUs in training large language models for AI. As described earlier, the '092 patent further describes that the offload processing modules can form a new switching plane or cloud fabric using their virtual switches for exchanging the training results between the offload processing modules and performing the second computing operations on the plurality of the offload processing modules in parallel for in-network computing of data-intensive workloads.

134. The '092 Patent solves a technological problem with server architectures for processing data. For example, the '092 Patent explains that "[e]nterprises store and process their large amounts of data in a variety of ways." '092 Patent, 1:31–32. One manner involves structured data (e.g., data stored in relational databases); "[h]owever, it is estimated that such formatted structured data represents only a tiny fraction of an enterprise's stored data." *Id.*, 1:32–41. "Organizations are becoming increasingly aware that substantial information and knowledge resides in unstructured data (i.e., 'Big Data') repositories." *Id.*, 1:41–44. But as the '092 Patent explains, "conventional platforms that are currently being used to handle structured and unstructured data can substantially differ in their architecture." *Id.*, 1:47–49. Accordingly, the '092 Patent recognized that "[a]n architecture that supports both structured and unstructured queries can better handle current and emerging Big Data applications." *Id.*, 1:55–57.

135. To address this issue, the '092 Patent discloses improved server architectures for processing data-intensive computing operations in parallel by utilizing offload processing

modules, or DPUs, which the ’092 Patent in one embodiment refers to as Xocket In-Line Memory Modules (XIMMs), to execute in-network computing operations in a new switching plane or cloud fabric. As the ’092 Patent explains, “[d]ata processing and analytics for enterprise server or cloud based data systems, including both structured or unstructured data, can be efficiently implemented on offload processing modules.” *Id.*, 2:51–54.

136. Using one or more offload processing modules, or DPUs, “it is possible to execute lightweight data processing tasks without intervention from a main server processor.” *Id.*, 2:56–61. In addition, “XIMM modules have high efficiency context switching, high parallelism, and can efficiently process large data sets. Such systems as a whole are able to handle large database searching at a very low power when compared to traditional high power ‘brawny’ server cores.” *Id.*, 2:61–66. Furthermore, “[a]dvantageously, by accelerating implementation of MapReduce or similar algorithms on unstructured data . . ., a XIMM based architecture capable of partitioning tasks is able to greatly improve data analytic performance.” *Id.*, 2:66–3:4.

137. For example, Claim 1 of the ’092 Patent is directed to:

> 1. A distributed computing architecture for executing at least first and second computing operations executed in parallel on a set of data, the architecture comprising:
>
> a plurality of servers, including first servers that each include
>
> > at least one central processing unit (CPU), and
> >
> > at least one offload processing module coupled to the at least one CPU by a bus, each offload processing module including a plurality of computation elements, the computation elements configured to
> >
> > > operate as a virtual switch, and
> > >
> > > execute the second computing operations on first processed data to generate second processed data; wherein

> the virtual switches form a switch fabric for exchanging data between the offload processing modules,
>
> the first computing operations generate the first processed data and are not executed by the offload processing modules, and
>
> the second computing operations are executed on a plurality of the offload processing modules in parallel.

138. NVIDIA is not licensed to the ’092 Patent.

139. Microsoft is not licensed to the ’092 Patent.

(4) The ’640 Patent – DPU In-Network Computing

140. U.S. Patent No. 9,436,640 (“the ’640 Patent”) is entitled “Full Bandwidth Packet Handling With Server Systems Including Offload Processors” for in-network computing operations, including in sorting, organizing, and reducing/combining data-intensive workloads in the new switching plane or new cloud fabric of the ’297 and ’161 Patents, also known to those of skill in the art as map/reduce operations. The ’640 Patent duly and legally issued on September 6, 2016, from U.S. Patent Application No. 13/931,910, filed on June 29, 2013.

141. The ’640 Patent claims priority from U.S. Provisional Application Nos. 61/753,892; 61/753,895; 61/753,899; 61/753,901; 61/753,903; 61/753,904; 61/753,906; 61/753,907; and 61/753,910, all filed on January 17, 2013. The ’640 Patent is entitled to the benefit of these earlier filed applications.

142. Xockets is the current owner of all rights, title, and interest in and to the ’640 Patent, including the right to sue for past damages.

143. A true and correct copy of the ’640 Patent is attached hereto as **Exhibit 7** and is incorporated by reference herein.

144. The ’640 Patent generally relates to systems, hardware, and methods for cloud data centers to create a rack server system with offload processor modules, or DPUs, for in-network

reduction/combining of data-intensive workloads using map/reduce data processing, which is critical in training large language models for AI. The rack server system includes a plurality of servers arranged in a rack, and a plurality of offload processor modules, or DPUs, supported on the servers. Each offload processor module has multiple offload processors that function as programmable hardware accelerators and an input-output (IO) port. The offload processor modules are connected directly to each other through their respective IO ports to form a midplane switch fabric for cloud offload of data-intensive workloads from server processors that can overcome the limitations associated with the existing cloud network and server systems. The offload processor modules are configured to execute map and reduce steps, thereby accelerating map/reduce data processing including collective communication of training results in training large language models for AI as well as in-network reduction/combining of the training results.

145. Like the ’161 Patent, the ’640 Patent solves a technological problem with server systems, including systems used in data centers and for data processing. For example, the ’640 Patent explains that “[e]fficient managing of network packet flow and processing is critical for high performance,” and “[s]ubstantial improvements in network service would be made possible by systems that can flexibly process a data flow, recognize or characterize patterns in the data flow, and improve routing and processing decisions for the data.” ’640 Patent, 1:25–34. “Unfortunately, the tree-like server connection topology often used in conventional data centers can be prone to traffic slowdowns and computational bottlenecks.” *Id.*, 1:35–37. “Typically, all the servers in such data centers communication with each other through higher level Ethernet-type switches, such as Top-Of-Rack (TOR) switches.” *Id.*, 1:37–40. However, as the ’640 Patent explains, “[f]low of all the traffic through such TOR switches leads to congestion results in increased network latency, particularly during the periods of high usage.” *Id.*, 1:40–43.

146. Improving upon the prior art, the ’640 Patent discloses and claims systems, hardware, and methods relating to rack server systems for packet processing. For example, “to prevent data transfer bottlenecks through TOR switches, and/or to improve[] system performance,” the ’640 Patent discloses systems in which “direct inter-rack and/or intra-rack communication can be enabled by offload processor modules included in the servers,” bypassing the existing cloud network. *Id.*, 4:20–24. “[S]uch data communication via offload processor modules can require less time and/or less processing power as compared to TOR switching via aggregation layer transfers. Accordingly, such data transfers can be executed in a more efficient manner than conventional systems.” *Id.*, 4:31–35. In addition, “[a]dvantageously, inter/intra-rack communications via offload server modules can also reduce the need for additional TOR switches and can be included to increase bandwidth and introduce redundancy, particularly since TOR switches may have to be periodically replaced to handle higher network speeds.” *Id.*, 4:36–41.

147. As further explained in the ’640 Patent, “[s]ervers equipped with offload processor modules, such as described herein and equivalent, can bypass a TOR switch through intelligent virtual switching of the offload processor modules associated with each server” and provide in-network computing operations for map/reduce data processing. *Id.*, 18:38–42.

148. For example, Claim 9 of the ’640 Patent is directed to:

> 9. A rack server system for a map/reduce data processing, comprising:
>
> a plurality of servers arranged in a rack,
>
> a plurality of offload processor modules supported on at least two of the servers, each offload processor module having an input-output (IO) port and multiple offload processors, a first offload processor module configured to execute map steps of the map/reduce data processing, and being connected directly to a second offload processor through their respective IO ports to define a midplane switch, and

> a top of rack (TOR) unit connected to each of the servers that does not transfer map/reduce data, wherein
>
> a second offload processor module is configured to execute reduce steps of the map/reduce data processing on data provided from the first offload processor module.

149. NVIDIA is not licensed to the '640 Patent.

150. Microsoft is not licensed to the '640 Patent.

II. BACKGROUND ON XOCKETS' CLOUD COMPUTING INVENTIONS

151. The advent of cloud computing has radically changed the computing industry. Instead of individual businesses running a relatively small number of general-purpose server platforms in-house, a relatively few number of specialists have arisen to run massive warehouses of computers. These specialists—cloud operators—could reap efficiencies of scale that individual owners could never reach. The cloud computing model has had profound economic, competitive, and technological implications for the industry—and society as a whole.

152. At first, these cloud operators built systems that looked like a traditional data center, just scaled up. Their systems revolved around server processors (Central Processing Units (CPUs), Graphical Processing Units (GPUs), and hybrids of these server processors). To meet the increased data demands of the growing cloud market, operators of these server processor-based cloud systems expected to rely on the performance increase of each new generation of server processors. These types of processors are optimized for complex programs that make lots of interdependent decisions and perform complicated math. Internally, they spend a tremendous amount of energy in identifying shortcuts through those interdependent decisions and include massive and power-hungry structures for number crunching.

153. And in fact, for the bulk of the history of the processor, the industry could count on steady performance increases in each new generation, resulting from the predictable density

improvement of transistors, known as Moore's Law. For a given power budget, performance improvement in processors was essentially guaranteed. Designers could add complexity to software and count on transistor performance to accelerate server processors to the point where they could handle it.

154. Dr. Dalal foresaw that Moore's Law would end, and its death would come at a most inopportune time, as he also anticipated that the data demands of cloud computing would exponentially increase. He turned out to be right on both scores. Transistor performance scaling through increased transistor density has in fact slowed and, for some metrics, has essentially stopped. Yet the amount of data traffic in the cloud has increased exponentially. We now live in the "Zettabyte Era." A zettabyte is the current largest digital unit of measurement. A zetta-stack of dollar bills would reach from the earth to the sun (93 million miles away) and back—700,000 times. According to Google's former CEO Eric Schmidt, from the beginning of humanity to the year 2003, an estimated 0.5% of a zettabyte was created.[25] In 2012, the year that Xockets filed its first provisional patent application, the amount of all digital data in the world first exceeded a zettabyte. Today, the volume of cloud traffic alone is estimated to be 50 zettabytes a year and growing. This is an almost inconceivable increase in data.

155. Anticipating the death of Moore's Law at a time when cloud data demands would exponentially increase, Dr. Dalal recognized that server processors would become a bottleneck, hitting walls of efficiency and performance. The conventional wisdom of adding more, ever-faster server processors would not solve the problems that the cloud computing platforms of the future would face (and help cause). These were the wrong kind of processors for processing the data-

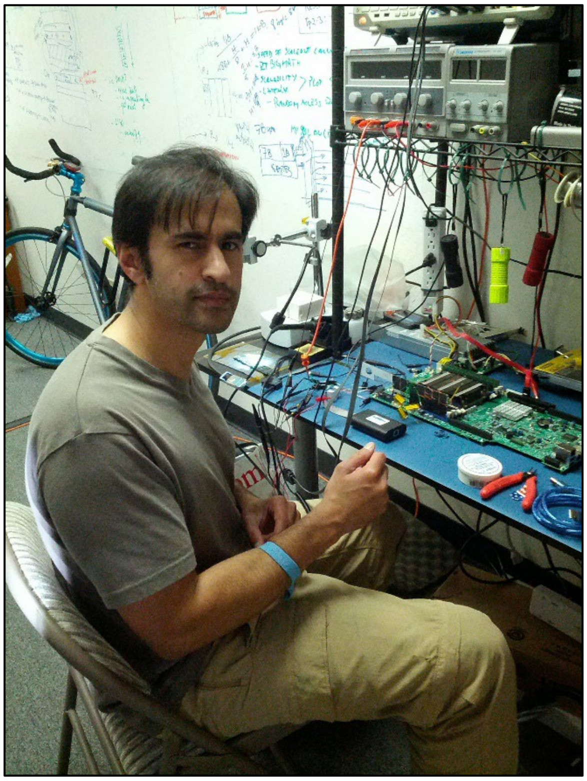
[25] Google Chief Eric Schmidt Keynote Speech at Guardian Activate 2010, Part 2, https://youtu.be/jcBPgEGA7Yk?si=WpIFgbu3Kwjr39Hw&t=227 (3:47–4:05).

intensive workloads required in cloud distributed computing. First, adding more and more server processors was a significant cost. Second, it did not address the root of the problem, which was that server processors were not designed to efficiently handle the data intensive infrastructure services that are required when massive amounts of data and processing are being pushed to the cloud. No matter how many server processors were added, bottlenecks from the data intensive infrastructure services would continue to arise, interrupting the server processors from running what they were designed to run: revenue-producing cloud applications, including training large language models for AI.

156. Dr. Dalal founded Xockets in 2012. He raised seed funding and then hired a team of network infrastructure engineers to help him develop his technology.

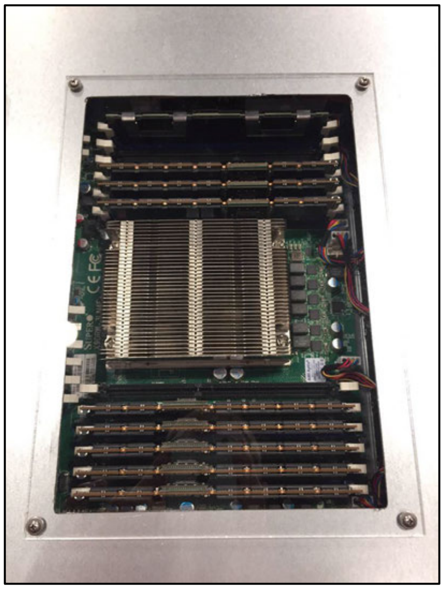
157. In May 2012, the first Xockets provisional patent application was filed. Additional provisional patent applications were filed in January 2013 and March 2014. These provisional patent applications disclosed, among other things, Xockets' DPU-based cloud architecture which ***offloads*** from CPUs, GPUs, and/or other host processors in servers ***and accelerates*** the data plane and/or control plane of cloud infrastructure services ***independent*** of server processors. For example, in Xockets' patented DPU-based cloud architecture, packet processing operations of key infrastructure services in cloud distributed computing may be offloaded from server processors to DPUs, including cloud-specific security, networking, and storage infrastructure services. As another example, brokering and accelerating communications between server processors (such as GPUs) in training large language models for implementing machine learning/artificial intelligence in cloud applications, referred to as cloud "ML/AI collective communications," and related computational operations, for sorting, organizing, and reducing/combining training results, may also be offloaded from server processors to DPUs.

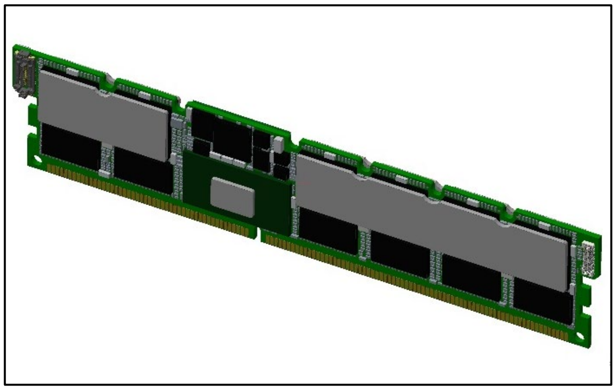
158. Dr. Dalal and his team at Xockets proceeded to design, develop, and build the world's first DPU for cloud offload processing.



159. In late September of 2015, Xockets demonstrated its DPU computing and switching architectures implemented in its StreamSwitch product at Strata, the industry's premier big data and network technology conference.

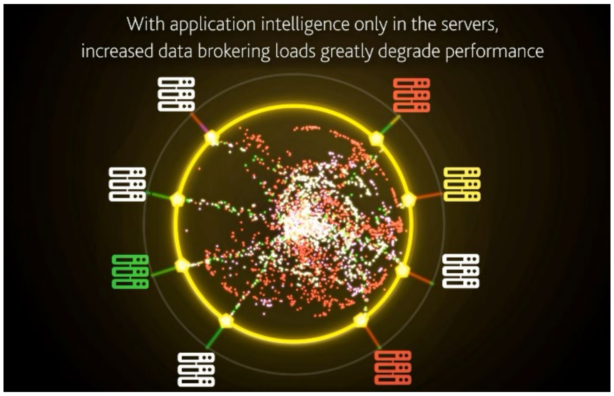
XOCKETS + SMART Modular Technologies
A Revolution in Scale-Out Architecture
StreamSwitch
Transforms Big Data Performance
Active Application Processing embedded into the network
Job Completion Time reduced 100X
No change to in-place hardware or software infrastructure
Reduces Big Data cluster TCO up to 40%
Active Application Network
Solve your hardest computational problems in real time with no change to your existing cluster and software tools
100X higher cluster efficiency for Hadoop
StreamSwitch
243mW / Gbps
XOCKETS + SMART Modular Technologies
A Revolution in Scale-Out Architecture

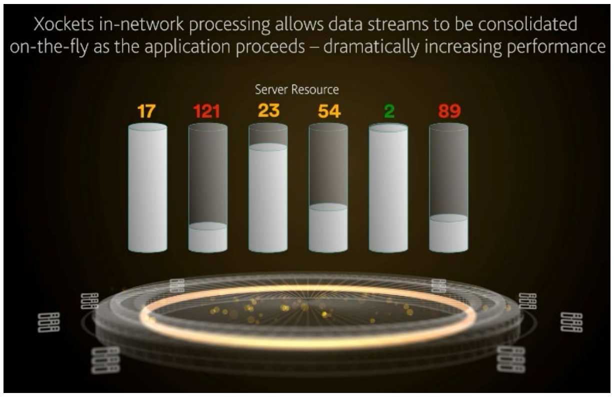


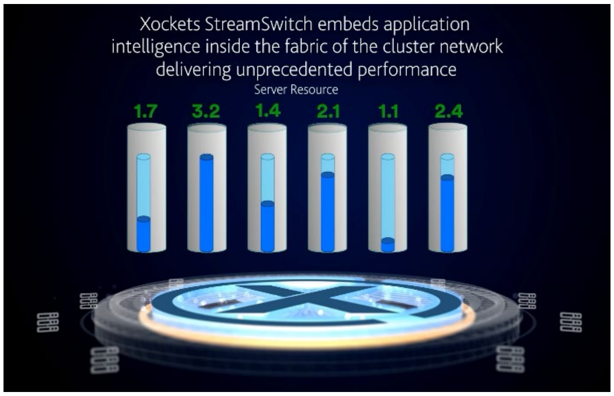


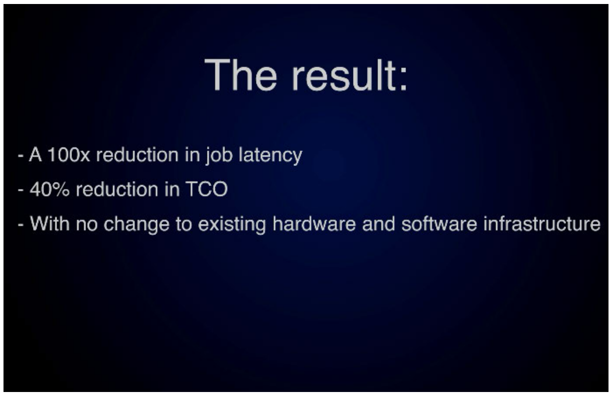
160. As Xockets recognized and showed, distributed computing in a cloud requires computing operations on billions of packets per second flowing between servers, including for cloud-specific security, networking, and storage, and the brokering of collective communications across server processors and in-network reduction of the data for further processing. Xockets'

presentations demonstrated the need for programmable hardware acceleration to scale distributed computing in modern cloud workloads, including cloud-specific "security," "flow/packet-based services," "big data," "encryption/decryption," and "machine learning" workloads for AI.

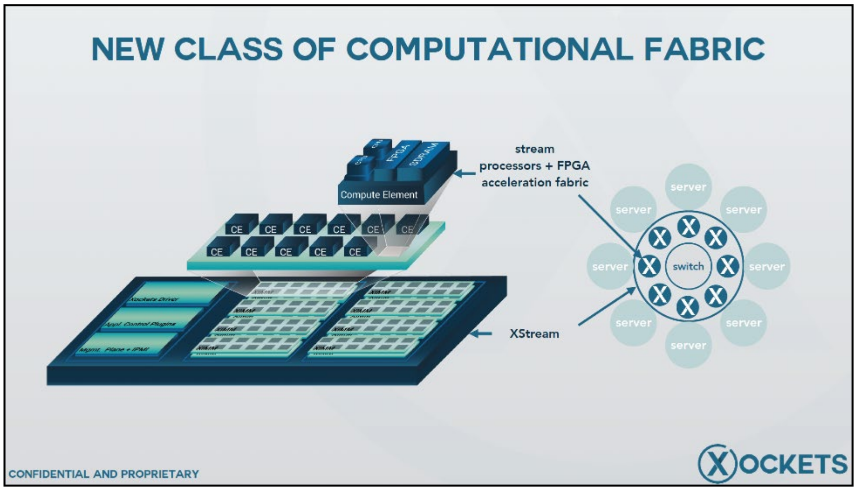

With application intelligence only in the servers, increased data brokering loads greatly degrade performance

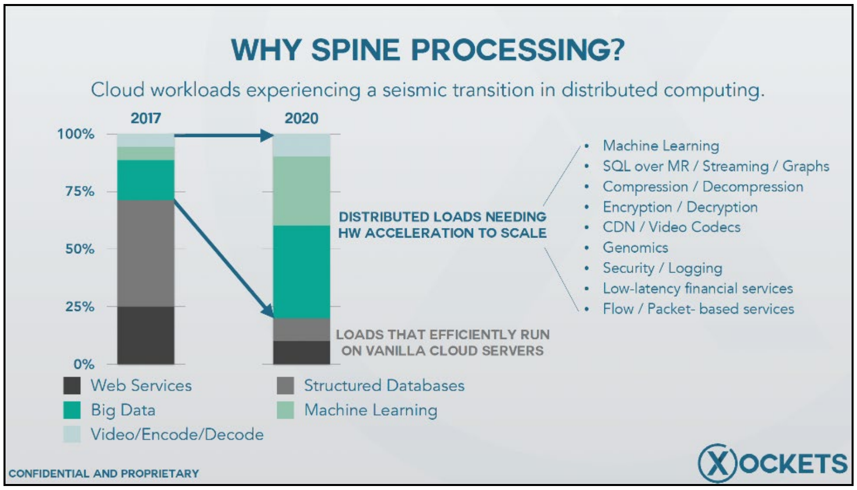

Xockets in-network processing allows data streams to be consolidated on-the-fly as the application proceeds – dramatically increasing performance


Xockets StreamSwitch embeds application intelligence inside the fabric of the cluster network delivering unprecedented performance


The result:

- A 100x reduction in job latency
- 40% reduction in TCO
- With no change to existing hardware and software infrastructure


NEW CLASS OF COMPUTATIONAL FABRIC



WHY SPINE PROCESSING?

Cloud workloads experiencing a seismic transition in distributed computing.

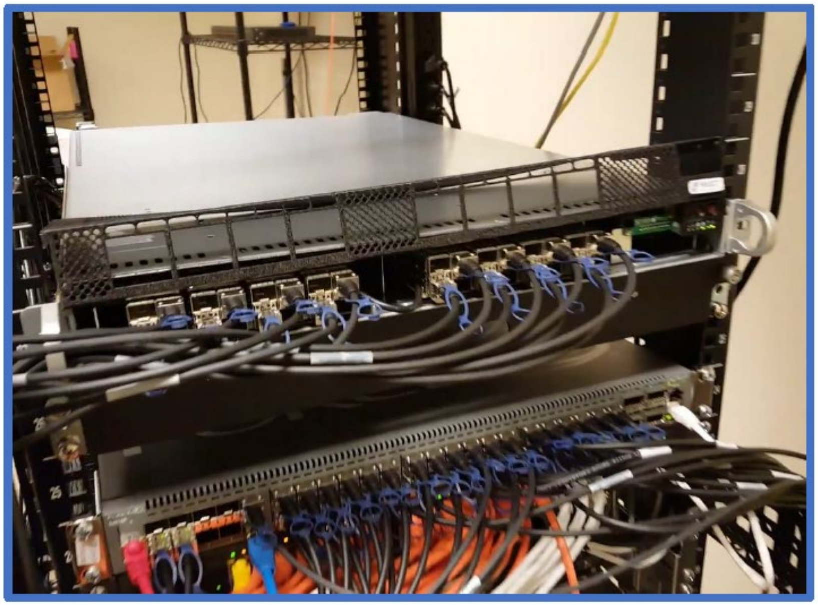
- Machine Learning
- SQL over MR / Streaming / Graphs
- Compression / Decompression
- Encryption / Decryption
- CDN / Video Codecs
- Genomics
- Security / Logging
- Low-latency financial services
- Flow / Packet- based services

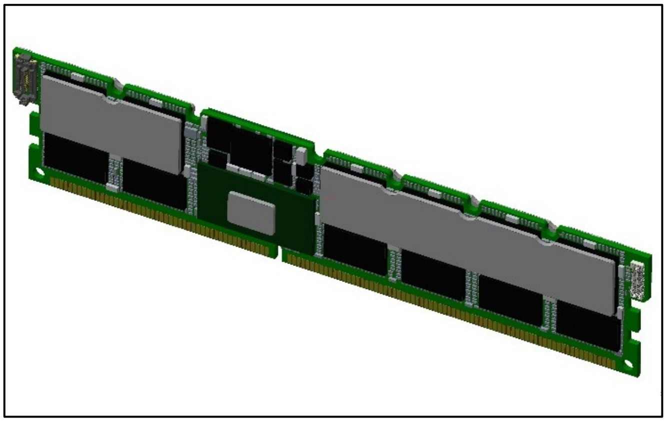
CONFIDENTIAL AND PROPRIETARY

XOCKETS

161. The StreamSwitch included 14 Xockets DPUs, one for each of 14 servers in a standard server rack. The DPUs were branded "Xocket In-line Memory Modules (XIMMs)" in this top of rack embodiment and provided "bump-in-the-wire acceleration" at "line rate" (i.e., network speed, then 10 Gbps) at the boundary between the network and each server processor and were connected at the network server interface through to the server's system bus.

162. This DPU embodiment included a ***Xockets virtual switch computing architecture*** for identifying and classifying packet flows and connecting together programmable logic pipelines of hardware accelerators for performing computational operations on the packet data for bump-in-the-wire offload and hardware acceleration of the cloud data plane and/or control plane independent of server processors. The photographs below show a Xockets StreamSwitch at Top of Rack, and the inside of a StreamSwitch, depicting the Xockets DPUs:





163. The Xockets team achieved up to ***1000x acceleration*** in the performance of Hadoop big data applications running on the StreamSwitch compared to host CPUs performing the same cloud-specific data plane infrastructure services. This performance benefit would result in (1) substantially higher revenues in a cloud data center as a result of freeing up server processors

for running more revenue-producing customer applications and services and by enabling accelerated computing performance in a cloud; and (2) substantially lower Total Cost of Ownership (TCO) for servers deployed in a cloud data center, resulting from a reduction of both capital expenses (fewer servers/CPUs for the same performance) and operating expenses (lower energy costs for the same performance).

164. After years of heading in the wrong direction, and following the publication of Xockets' patent applications and its public demonstrations of its pioneering new DPU-based architecture, NVIDIA abandoned conventional server processor-centric approaches and embraced implementing Xockets' patented architecture by utilizing DPUs for cloud offload processing. Xockets' patented DPU-based cloud architecture is now [redacted]

III. NVIDIA'S USE OF XOCKETS' PATENTED TECHNOLOGY

165. NVIDIA began its business in the graphics market. However, users soon realized that NVIDIA's graphics processors could be used as very high-performance server processors to perform massive technical number crunching. This led NVIDIA to begin selling more expensive processors into more lucrative markets such as cryptocurrency.

166. However, NVIDIA continued to operate as a component vendor for many years. In 2012, NVIDIA graphics processors were added to the Department of Energy's Titan supercomputer. But the complete system was designed and built by Cray, using AMD microprocessors and Cray-designed networking. This system was followed by two more supercomputers, Summit and Sierra, which were designed by IBM using NVIDIA GPUs and Mellanox-designed networking.

167. NVIDIA eventually realized that it needed to move from being a component vendor to being a system designer and provider. NVIDIA explained: "While computing demand is

surging, CPU performance advances are slowing as Moore's law has ended. This has led to the adoption of accelerated computing with NVIDIA GPUs and Mellanox's intelligent networking solutions. Datacenters in the future will be architected as giant compute engines with tens of thousands of compute nodes, designed holistically with their interconnects for optimal performance."[26] This solution NVIDIA hit upon was to take Xockets' patented technology.

168. NVIDIA uses Xockets' New Cloud Processor and New Cloud Fabric technology.

A. NVIDIA'S KNOWLEDGE OF THE XOCKETS PATENTS

169. NVIDIA has had actual notice of Xockets' New Cloud Processor and New Cloud Fabric Patents at least since February 2022.

170. NVIDIA is and was well aware of Xockets' breakthrough invention of DPU computing architecture and switching fabric as detailed herein, including the invention of a virtual switch for implementing programmable hardware acceleration in the network for cloud offload of data-intensive distributed computing operations independent of host CPUs/GPUs in servers and the invention of a switching fabric for connecting CPUs/GPUs independent of the existing cloud network for cloud offload of data-intensive distributed computing operations, including training large models for AI.

171. On January 27, 2022, Xockets' co-founder, Dr. Dalal, communicated with NVIDIA's Brad Genereaux (Global Lead, Healthcare Alliances) and asked for an introduction to "Nvidia legal IP Counsel" in order to discuss "some very strategic IP" that "Nvidia would be interested in acquiring." Dr. Dalal sought to present NVIDIA the opportunity to acquire exclusive rights to Xockets' patent portfolio. After making an internal inquiry, Mr. Genereaux ultimately connected Dr. Dalal with Gady Rosenfeld on February 4, 2022.

[26] https://nvidianews.nvidia.com/news/nvidia-to-acquire-mellanox-for-6-9-billion.

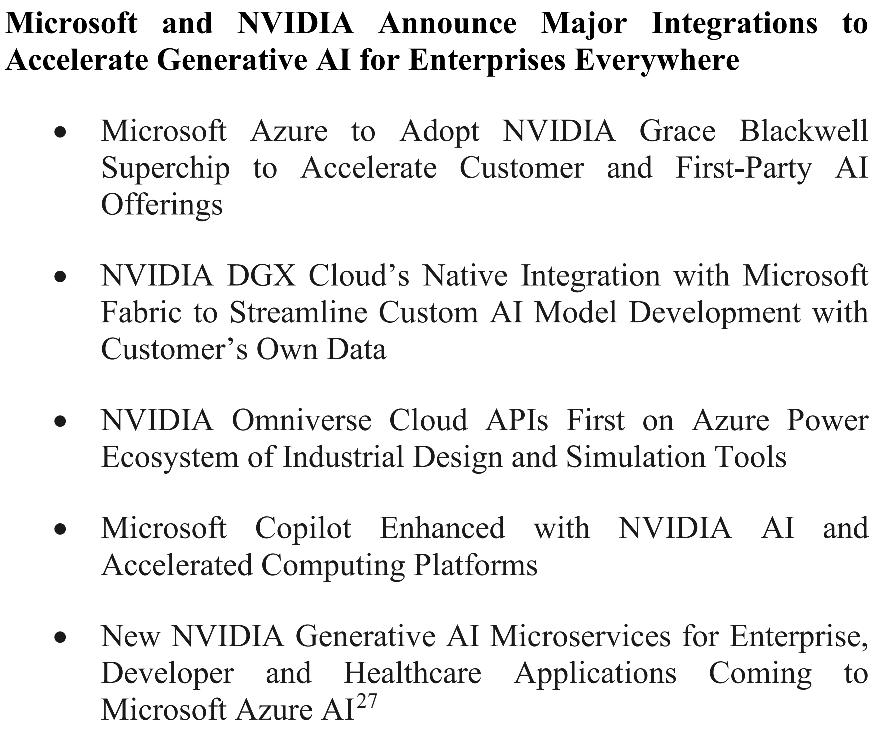
172. Mr. Rosenfeld was "leading the DPU segment in the NVIDIA field organization" at that time. Indeed, Mr. Rosenfeld's LinkedIn profile reflects that he has been NVIDIA's Vice President, DPU Business since July 2021 and remains in that role today. Dr. Dalal and Mr. Rosenfeld had a Teams meeting on February 10 to discuss Xockets and its IP. Dr. Dalal walked Mr. Rosenfeld through exemplary claim charts and explained the nature of Xockets' patented technology. Mr. Rosenfeld indicated during that meeting that the technology was "extremely interesting." Later that same day, Dr. Dalal emailed Mr. Rosenfeld sample claim charts and a list of Xockets' then-current patent list covering breakthrough DPU technologies essential to AI, which included the New Cloud Processor Patents (the '209, '924, and '350 Patents) and the New Cloud Fabric Patents (the '297, '161, '092, 'and '640 Patents).

173. Mr. Rosenfeld told Dr. Dalal that he would discuss Xockets' patent portfolio with NVIDIA's legal department and then follow up on next steps.

IV. MICROSOFT'S USE OF XOCKETS' TECHNOLOGY

174. Microsoft holds the dominant market position in GPU-enabled generative AI platforms via its agreements with leading generative AI model companies, including OpenAI, and its agreements with NVIDIA. Microsoft is in the process of creating and/or maintaining a monopoly in this field.

175. NVIDIA and Microsoft have formed a cartel to monopolize GPU-enabled generative artificial intelligence by controlling the equipment and platforms necessary to access this capability. For example, Microsoft and NVIDIA publicly tout that Microsoft is gaining first access to its GPU-enabled generative AI servers and that Microsoft is embedding NVIDIA technology into the GPU-enabled generative AI platform market it dominates. This creates a self-reinforcing cycle in which users who desire this capability have no choice but to use NVIDIA and Microsoft because of their dominant combined position.



Microsoft and NVIDIA Announce Major Integrations to Accelerate Generative AI for Enterprises Everywhere

- Microsoft Azure to Adopt NVIDIA Grace Blackwell Superchip to Accelerate Customer and First-Party AI Offerings
- NVIDIA DGX Cloud's Native Integration with Microsoft Fabric to Streamline Custom AI Model Development with Customer's Own Data
- NVIDIA Omniverse Cloud APIs First on Azure Power Ecosystem of Industrial Design and Simulation Tools
- Microsoft Copilot Enhanced with NVIDIA AI and Accelerated Computing Platforms
- New NVIDIA Generative AI Microservices for Enterprise, Developer and Healthcare Applications Coming to Microsoft Azure AI[27]

176. NVIDIA advertises that Microsoft Azure uses NVIDIA's "DGX Cloud" system.[28]

177. Both NVIDIA and Microsoft publicly promote Microsoft's use of the accused NVIDIA DPU-enabled systems that copy Xockets' technology. For example, NVIDIA advertises that "Microsoft Azure and NVIDIA are empowering enterprises to achieve new levels of innovation. With NVIDIA's ***full-stack accelerated computing platform*** combined with Microsoft's global-scale, simplified infrastructure management, enterprises can transform their businesses."[29]

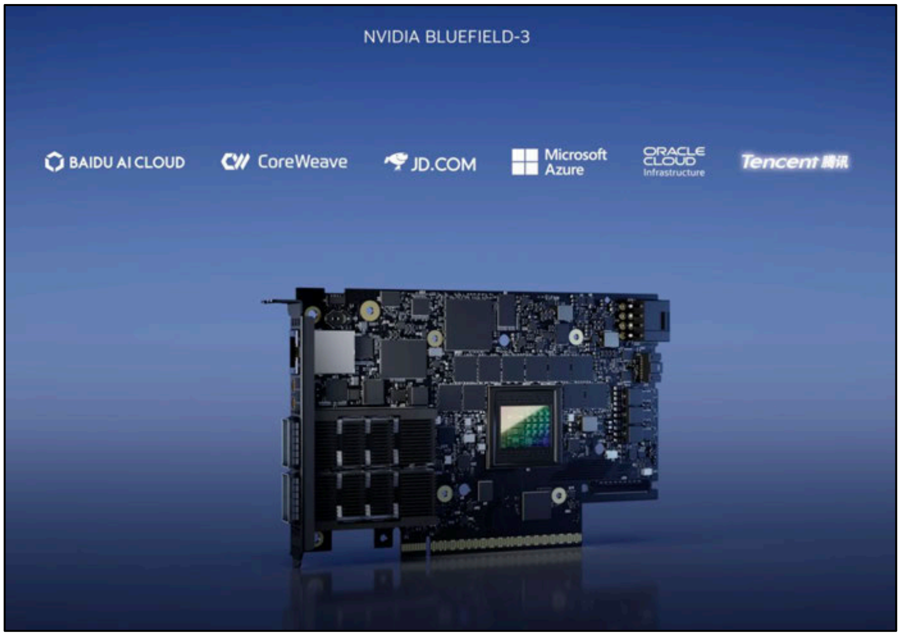
[27] https://nvidianews.nvidia.com/news/microsoft-nvidia-generative-ai-enterprises.

[28] https://www.nvidia.com/en-us/data-center/dgx-cloud; https://azuremarketplace.microsoft.com/en-us/marketplace/apps/nvidia.dgx-cloud?tab=Overview (referring to "NVIDIA DGX™ Cloud on Microsoft Azure").

[29] https://www.nvidia.com/en-us/data-center/dgx-cloud.

178. NVIDIA also advertises that it is "partnering with Microsoft to accelerate the development and deployment of generative AI across Microsoft Azure, Azure AI services, Microsoft Fabric, and Microsoft 365."[30]

179. NVIDIA also advertises that Microsoft Azure uses NVIDIA's BlueField DPUs[31]:



A. MICROSOFT'S PRAISE OF XOCKETS' PATENTED TECHNOLOGY

180. Microsoft has described building out its Azure infrastructure as "[t]he most important thing we've done over the last four years":

> "***The most important thing is what we've done over the last four years [since 2019] is to actually build out the core infrastructure on which OpenAI is built***. I mean, these large models, the training infrastructure and the [ML/AI] inference infrastructure doesn't look like just vanilla cloud, right? So ***we have had to essentially evolve***

[30] https://www.nvidia.com/en-us/events/microsoft-build.

[31] GTC 2023 Keynote with NVIDIA CEO Jensen Huang, https://www.youtube.com/watch?v=DiGB5uAYKAg&t=1884s (31:24–31:39).

> ***Azure [with NVIDIA] to be pretty specialized AI infrastructure. . . .***"[32]

181. In addition, Microsoft has boasted about the benefits the technology brings to Microsoft Azure:

> "***Together with NVIDIA, we are making the promise of AI real, helping drive new benefits and productivity gains for people and organizations everywhere***," said Satya Nadella, chairman and CEO, Microsoft. "***From bringing the GB200 Grace Blackwell processor to Azure, to new integrations between DGX Cloud and Microsoft Fabric***, the announcements we are making today will ensure customers have the most comprehensive platforms and tools across every layer of the Copilot stack, from silicon to software, to build their own breakthrough AI capability."[33]

182. Microsoft has emphasized the significance of its collaboration with NVIDIA in delivering "state-of-the-art AI capabilities for every enterprise on Microsoft Azure":

> "AI is fueling the next wave of automation across enterprises and industrial computing, enabling organizations to do more with less as they navigate economic uncertainties," said Scott Guthrie, executive vice president of the Cloud + AI Group at Microsoft. "***Our collaboration with NVIDIA unlocks the world's most scalable supercomputer platform, which delivers state-of-the-art AI capabilities for every enterprise on Microsoft Azure***."[34]

183. Microsoft has boasted that with NVIDIA, it is providing "the most powerful AI supercomputer" in the world to its customers:

> "The next wave of computing is being born, between next-generation immersive experiences and advanced foundational AI models, we see the emergence of a new computing platform," said Satya Nadella, chairman and CEO of Microsoft. "***Together with NVIDIA, we're focused on both building out services that bridge the digital and physical worlds to automate, simulate and predict***

[32] Why Microsoft's CEO is Ready to Take on Google with ChatGPT, https://www.youtube.com/watch?v=QinFy0RFDr8&t=163s (2:43–3:05).

[33] https://news.microsoft.com/2024/03/18/microsoft-and-nvidia-announce-major-integrations-to-accelerate-generative-ai-for-enterprises-everywhere.

[34] https://nvidianews.nvidia.com/news/nvidia-microsoft-accelerate-cloud-enterprise-ai.

> ***every business process, and bringing the most powerful AI supercomputer to customers globally.***"[35]

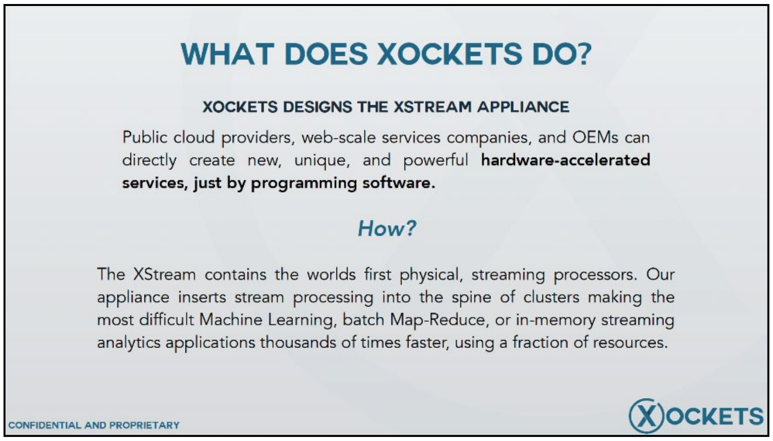
B. MICROSOFT'S KNOWLEDGE OF THE XOCKETS PATENTS

184. Microsoft is and was well aware of Xockets' breakthrough invention of DPU computing architecture and switching fabric as detailed herein, including the invention of a virtual switch for implementing programmable hardware acceleration in the network for cloud offload of data-intensive distributed computing operations independent of host CPUs/GPUs in servers and the invention of a switching fabric for connecting CPUs/GPUs independent of the existing cloud network for cloud offload of data-intensive distributed computing operations, including training large models for AI.

185. In 2015 at the most important big data technology conference in the world Xockets presented its technology.

186. Xockets and Microsoft began discussing the potential benefits to Microsoft of Xockets' technology in May 2016. In March 2017, after a large company expressed an interest in acquiring Xockets, Xockets' Dan Alvarez reached out to Microsoft's Ulrich Homann (Corporate Vice President, Cloud and AI) and Jim Brisimitzis (General Manager, Cloud Developer Relations) with a call for bids. The call for bids provided an overview of Xockets' technology and the fact that Xockets already had a large number of issued patents on the technology:

[35] https://nvidianews.nvidia.com/news/nvidia-and-microsoft-to-bring-the-industrial-metaverse-and-ai-to-hundreds-of-millions-of-enterprise-users-via-azure-cloud.



WHAT DOES XOCKETS DO?

XOCKETS DESIGNS THE XSTREAM APPLIANCE

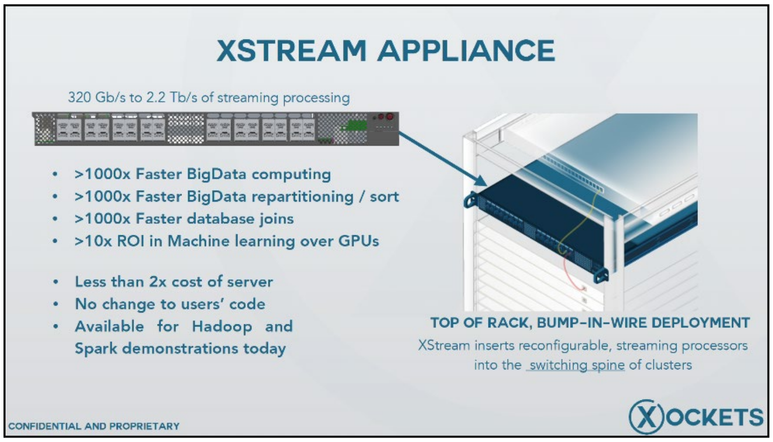
Public cloud providers, web-scale services companies, and OEMs can directly create new, unique, and powerful **hardware-accelerated services, just by programming software.**

How?

The XStream contains the worlds first physical, streaming processors. Our appliance inserts stream processing into the spine of clusters making the most difficult Machine Learning, batch Map-Reduce, or in-memory streaming analytics applications thousands of times faster, using a fraction of resources.

CONFIDENTIAL AND PROPRIETARY

XOCKETS



XSTREAM APPLIANCE

320 Gb/s to 2.2 Tb/s of streaming processing

- >1000x Faster BigData computing
- >1000x Faster BigData repartitioning / sort
- >1000x Faster database joins
- >10x ROI in Machine learning over GPUs

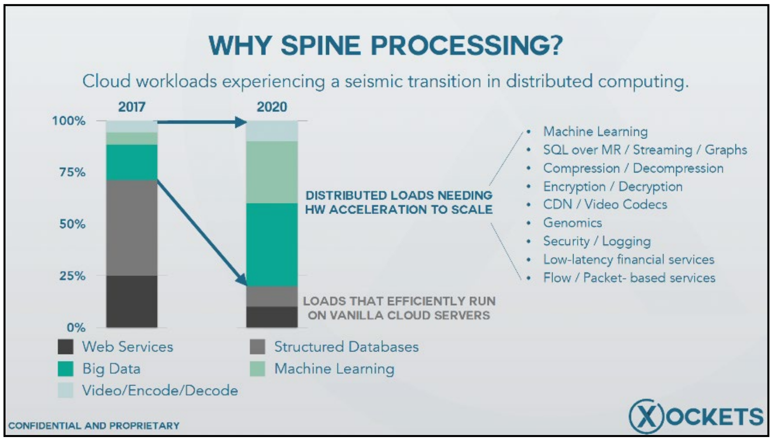
- Less than 2x cost of server
- No change to users' code
- Available for Hadoop and Spark demonstrations today

TOP OF RACK, BUMP-IN-WIRE DEPLOYMENT

XStream inserts reconfigurable, streaming processors into the switching spine of clusters

CONFIDENTIAL AND PROPRIETARY

XOCKETS



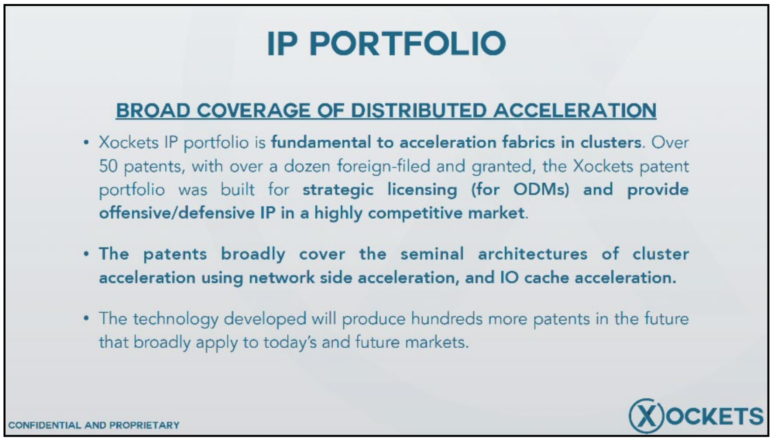
WHY SPINE PROCESSING?

Cloud workloads experiencing a seismic transition in distributed computing.

- Machine Learning
- SQL over MR / Streaming / Graphs
- Compression / Decompression
- Encryption / Decryption
- CDN / Video Codecs
- Genomics
- Security / Logging
- Low-latency financial services
- Flow / Packet- based services

CONFIDENTIAL AND PROPRIETARY

XOCKETS



IP PORTFOLIO

BROAD COVERAGE OF DISTRIBUTED ACCELERATION

- Xockets IP portfolio is **fundamental to acceleration fabrics in clusters**. Over 50 patents, with over a dozen foreign-filed and granted, the Xockets patent portfolio was built for **strategic licensing (for ODMs) and provide offensive/defensive IP in a highly competitive market**.
- **The patents broadly cover the seminal architectures of cluster acceleration using network side acceleration, and IO cache acceleration.**
- The technology developed will produce hundreds more patents in the future that broadly apply to today's and future markets.

CONFIDENTIAL AND PROPRIETARY

XOCKETS

Exhibit 8: "Xockets in a Nutshell" from Microsoft emails, at 2–4, 8.

187. In response, Mr. Homann responded that the "concept resonates and the team would like to understand in more depth." Mr. Homann directed Xockets to interface with Saurabh Kulkarni (Director of Engineering, Cloud and AI System Technologies) and Kushagra Vaid (VP and Distinguished Engineer, Azure Infrastructure). Ultimately, Dr. Dalal had a discussion with Mr. Kulkarni and Tanj Bennett (Partner SDE) on March 22, 2017, so they could "get a technical overview of key Xockets technologies in the hardware acceleration space." Thereafter, Mr. Kulkarni informed Dr. Dalal that he was reaching out to folks from Microsoft's "big data and machine learning teams" in order to make an introduction. As discussed further below, instead of further engaging with Xockets regarding its technology, Microsoft just took Xockets' technology without paying for it. And when Xockets subsequently approached Microsoft about taking a license, Microsoft formed a buyers' cartel with NVIDIA whereby both of them agreed not to compete for the license or purchase of Xockets' technology, but instead to negotiate only through RPX in order to obtain a price below what would have been obtained under normal, non-collusive market conditions.

V. REPRESENTATIVE BENEFITS OF XOCKETS' PATENTED TECHNOLOGY

188. Implementing Xockets' patented inventions through DPUs for cloud offload processing provides multiple benefits, including: Total Cost of Ownership ("TCO") Savings and Accelerated Performance.

189. **Increased TCO Savings:** With respect to TCO Savings—as described above, data centers are under increasing pressure to keep operating costs down. Innovations which provide the opportunity to purchase less hardware equipment (such as CPUs or GPUs), and use less power (thus saving on power consumption costs), are hugely valuable to data centers. Xockets' inventions provide these exact benefits.

190. DPUs for cloud offload processing enhance server efficiency by offloading data intensive infrastructure tasks involved in managing the flows of packets in a cloud data center, thereby freeing up valuable CPU or GPU cycles. Without Xockets' patented inventions, these infrastructure tasks were previously performed by the server processors, such as CPUs or GPUs. By offloading these tasks to DPUs, freeing up CPU or GPU cycles, cloud data centers can do more processing with less hardware, maximizing their return on investment. As one example, NVIDIA estimates that a single DPU can replace 300 CPUs.[36]

191. Cloud operators can exchange these TCO savings for denser, higher performance data centers that can produce higher revenues and profits by running more customer applications and services and accelerating their performance.

192. NVIDIA has performed studies corroborating these benefits. For example, in a November 2022 White Paper titled "DPU Power Efficiency," attached hereto as **Exhibit 9**,

[36] https://nvidianews.nvidia.com/news/nvidia-extends-data-center-infrastructure-processing-roadmap-with-bluefield-3.

NVIDIA estimated that DPUs for cloud offload processing can "reduce server power consumption up to 30%. . . . plus additional savings in cooling, power deliver, rack space, and server capital costs." **Exhibit 9** at 4. NVIDIA also stressed the importance of power savings, explaining "[n]ow that most data centers can be brought online rapidly and offer high levels of availability and compute density, improving power consumption and reducing associated power costs have become top goals both for optimizing existing data centers and designing new ones." *Id.*

193. NVIDIA performed several tests which show that offloading different categories of tasks result in significant TCO Savings. As one specific example, NVIDIA found that the 3-year TCO savings from offloading ***only*** IPsec encryption/decryption to a BlueField DPU (such as used in Cloud VPN network overlay services), was approximately ***$3,207*** per server in 3-year TCO savings for ***cloud security offload*** (based on a savings of $26.3 million across 8,200 servers, as shown in the table below). *Id.* at 21, Table 7. NVIDIA explained: "We see significant two-way savings from the offload and acceleration capabilities of the BlueField DPU. The offload frees up CPU cores allowing fewer servers to be deployed, reducing CapEx. The lower number of servers and lower per-server power consumption combine to reduce OpEx substantially. The result is a substantial savings of $26M over three years in a large data center with 10,000 servers." *Id.*

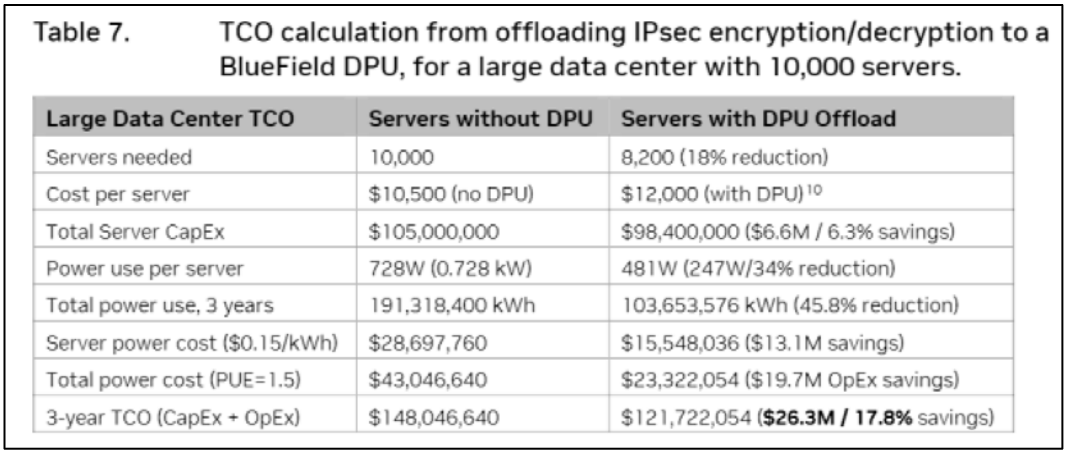


Table 7. TCO calculation from offloading IPsec encryption/decryption to a BlueField DPU, for a large data center with 10,000 servers.

Large Data Center TCO	Servers without DPU	Servers with DPU Offload
Servers needed	10,000	8,200 (18% reduction)
Cost per server	$10,500 (no DPU)	$12,000 (with DPU)[10]
Total Server CapEx	$105,000,000	$98,400,000 ($6.6M / 6.3% savings)
Power use per server	728W (0.728 kW)	481W (247W/34% reduction)
Total power use, 3 years	191,318,400 kWh	103,653,576 kWh (45.8% reduction)
Server power cost ($0.15/kWh)	$28,697,760	$15,548,036 ($13.1M savings)
Total power cost (PUE=1.5)	$43,046,640	$23,322,054 ($19.7M OpEx savings)
3-year TCO (CapEx + OpEx)	$148,046,640	$121,722,054 (**$26.3M / 17.8%** savings)

194. Significantly, each of these TCO savings were calculated based on the offload of just security and collective communication offload. In reality, as described above, DPUs allow for the offload of other cloud virtualization, network, storage, and security tasks. NVIDIA CEO Jensen Huang explained that these tasks in total "***can consume nearly half of the data center's CPU cores and associated power***."[37]

195. Huang confirmed that offloading all data-intensive cloud infrastructure services to DPUs can reduce approximately 50% of a cloud data center's power usage (OpEx) and server requirements (CapEx). By using NVIDIA's own method of calculating TCO Savings, the calculated 3-year TCO Savings of ***$15,457 per server*** in ***TCO savings*** (based on a savings of approximately $77.2 million across an estimated 5,000 servers, using NVIDIA's methodology in Table 7 above):

Large data center TCO	**Servers without DPU**	**Servers with DPU offload**
Servers needed	10,000	5,000
Cost per server	$10,500	$12,000
Total server CapEx	$105,000,000	$60,000,000
Power use per server	728 Watts (W)	364 Watts (W)
Total power use, 3 years	191,318,400 kWh	47,829,600 kWh
Server power cost ($0.15/kWh)	$28,697,760	$7,174,440
Total power cost (PUE = 1.5)	$43,046,640	$10,761,660
3-year TCO (CapEx + OpEx)	$148,046,640	$70,761,660
***3-year TCO Savings* ($/server)**	**-**	***$77,284,980 ($15,457/server)***
3-year TCO Savings (%)	-	52.2%

196. Xockets anticipates that the actual cost savings will be much greater.

197. NVIDIA has explained that by using DPUs for offloading cloud processing "[t]here will typically also be additional savings from the ability to run more revenue-generating workloads

[37] GTC 2023 Keynote with NVIDIA CEO Jensen Huang, https://www.youtube.com/watch?v=DiGB5uAYKAg&t=1836s (30:36–49).

as well on each server thanks to the CPU cycles freed up by the networking offload. Deploying DPU offloads in servers usually allows each server to perform more work (more connections, more virtual machines, more users, etc.). This results in a large CapEx savings because fewer servers are needed, as well as a significant OpEx savings because fewer servers consume less power, floor space, and other data center resources (cooling, power distribution, management)." **Exhibit 9** at 20.

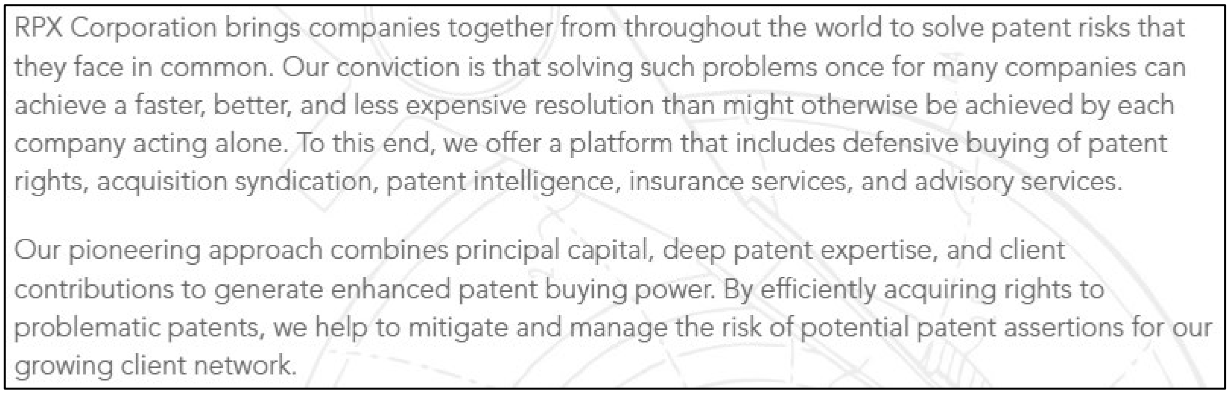
198. **Accelerated Performance:** The use of DPUs for cloud offload processing also accelerates the performance of cloud applications by enabling additional compute cycles, which results in reducing latency and an enhanced end user experience. The improved application responsiveness and reliability have a direct impact on customer satisfaction, user engagement, and higher transaction volumes and prices, all of which contribute to increased prices and revenue.

199. For example, on the low end, NVIDIA has estimated that "virtualization, networking, storage, security, management, and provisioning" can "consume up to 30% of the processor cycles." *Id*. at 8. By using DPUs to free up those cycles, performance is improved and server processor cores are able to run the types of applications they do best. *Id*.

200. Each offload use therefore results in accelerated application performance and reduced latency which further leads to enhanced customer satisfaction, increased user engagement, and higher market share/prices, which in turn leads to increased revenues. On information and belief, NVIDIA accordingly charges higher prices for products that allow for increased acceleration of the performance of cloud applications. On information and belief, this accelerated performance is estimated to provide public cloud providers like Microsoft at least ***$30,000 per server*** in ***increased revenue*** over the 3-year lifespan of a server in cloud data centers.

VI. RPX'S BUSINESS

201. RPX was founded in 2008 and has more than 450 members, including NVIDIA and Microsoft. RPX's website explains:



RPX Corporation brings companies together from throughout the world to solve patent risks that they face in common. Our conviction is that solving such problems once for many companies can achieve a faster, better, and less expensive resolution than might otherwise be achieved by each company acting alone. To this end, we offer a platform that includes defensive buying of patent rights, acquisition syndication, patent intelligence, insurance services, and advisory services.

Our pioneering approach combines principal capital, deep patent expertise, and client contributions to generate enhanced patent buying power. By efficiently acquiring rights to problematic patents, we help to mitigate and manage the risk of potential patent assertions for our growing client network.

[38]

202. RPX previously touted on its website (language that has since been removed) that "[i]n effect, RPX can buy 'wholesale' on behalf of our client network, while our clients otherwise would pay 'retail' if transacting on their own." **Exhibit 10**. The RPX website also previously advertised that "RPX is often able to achieve 'wholesale' pricing terms, where we can acquire rights for our members at significantly reduced cost relative to what the NPE might charge an individual company on its own. RPX believes we have saved our members tens of millions of dollars through these wholesale-priced transactions." **Exhibit 11**. Despite RPX having removed the language in an effort to hide its anticompetitive behavior, RPX's business practices remain the same today.

203. RPX's most recent 10-K filing with the SEC in 2018 explains its mission of interjecting itself as the "essential intermediary" between patent owners and RPX's members:

[38] https://www.rpxcorp.com/about.

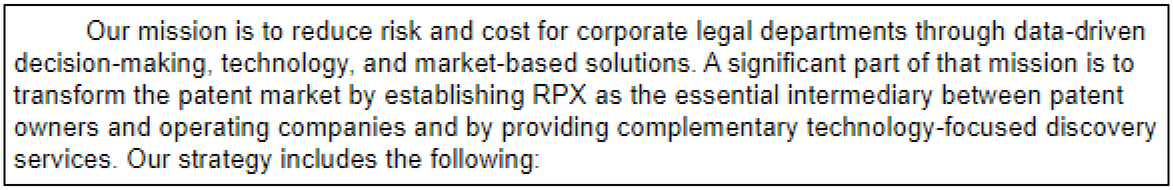
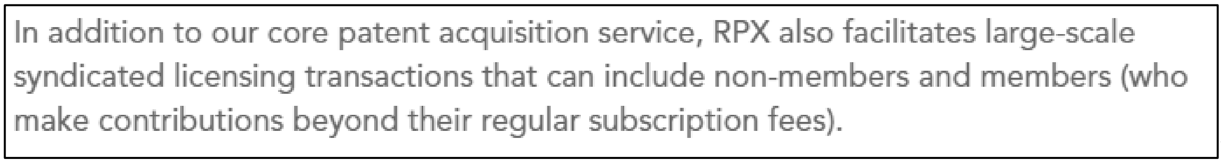

Our mission is to reduce risk and cost for corporate legal departments through data-driven decision-making, technology, and market-based solutions. A significant part of that mission is to transform the patent market by establishing RPX as the essential intermediary between patent owners and operating companies and by providing complementary technology-focused discovery services. Our strategy includes the following:

Exhibit 12, at 6. RPX's business practices remain the same today.

204. RPX's co-founder and former CEO, John Amster, has publicized RPX's mission, stating that "[w]e think there can be a clearinghouse in this market that can be really quite big and efficient. If every company just decided, 'We're going to have a line item in our budget for patents and patent risks, and that line item is going to be the RPX rate card'—i.e. RPX's subscription rate[.]"

205. Indeed, RPX's website (which refers to the "RPX Network" as the "world's leading defensive patent acquisition network") touts how RPX's application of "capital to acquire patents rights" leads to "far less cost" for its members, that "[t]here's safety in numbers" and "huge cost savings, too."[39]

206. RPX's website also explains how it collaborates with its members and non-members to create anti-competitive buyers' cartels, what it euphemistically calls "syndicated licensing transactions":


In addition to our core patent acquisition service, RPX also facilitates large-scale syndicated licensing transactions that can include non-members and members (who make contributions beyond their regular subscription fees).

[40]

207. RPX previously highlighted the benefits of these syndicated transactions for its clients on its website, in language that has since been removed, stating that "[o]ur clients see

[39] www.rpxcorp.com/solutions/rpx-network.

[40] http://ir.rpxcorp.com.

distinct advantages of syndicated purchasing through RPX, as we are uniquely situated to structure transactions that are ultimately less costly and deliver more value to participating clients than if any attempted individual licensing or unilateral purchasing of the portfolios." **Exhibit 10**.

208. RPX has openly acknowledged in its SEC filings that its practices may be illegal and violate competition and antitrust laws, admitting that "[i]t is possible that courts or other governmental authorities will interpret existing laws regulating [] competition and antitrust practices [] in a manner that is inconsistent with our business practices."[41]

VII. NVIDIA AND MICROSOFT RESPOND TO XOCKETS' 2024 FUNDRAISING EFFORTS BY CREATING A BUYERS' CARTEL

209. In early 2024, Xockets engaged in a process to sell or license its technology [redacted]. As part of the effort, NVIDIA and Microsoft were approached about the Xockets technology. Specific to NVIDIA, on March 27, 2024, Xockets' representative emailed NVIDIA's Vishal Bhagwati (Head of Corporate Development), Timothy Teter (Executive Vice President and General Counsel), David Shannon (EVP, Chief Administrative Officer and Secretary), and Rich Domingo (Director of Intellectual Property) Xockets' information and a proposed NDA. Xockets' representative separately followed up with Mr. Domingo on April 2 and 9 and June 5, and Gady Rosenfeld (NVIDIA's Vice President, DPU Business) on May 2. On April 30, Xockets' representative also sent a follow up email to his original March 27 email to Messrs. Bhagwati, Teter, Shannon, and Domingo. Around the time of these emails or shortly thereafter, NVIDIA interacted with RPX to form a conspiracy with Microsoft to create a buyers' cartel by refusing to negotiate individually with Xockets and instead only negotiating through RPX.

[41] RPX Corporation, S.E.C. Registration Statement (Form S-1), at 17 (Jan. 21, 2011), available at https://www.sec.gov/Archives/edgar/data/1509432/000119312511012087/ds1.htm.

210. Specific to Microsoft, on March 27, 2024, Xockets' representative emailed Microsoft's Christopher Young (Executive Vice President Business Development), Michael Wetter (Corporate Vice President, Corporate Development), and Nicholas Kim (Senior Corporate Counsel, IP Litigation) the teaser and a proposed NDA. He forwarded that email to Microsoft's Steve Bathiche day later and Brad Smith (President) on April 2. On April 30, Xockets' representative also followed up the original email with Messrs. Young, Wetter, and Kim, and separately with Mr. Smith. Around the time of these emails or shortly thereafter, Microsoft interacted with RPX to form a conspiracy with NVIDIA to create a buyers' cartel by refusing to negotiate individually with Xockets and instead only negotiating through RPX.

211. In May 2024, RPX's CEO, Dan McCurdy, contacted Xockets' representative to have a call and set up a subsequent dinner meeting. During the conversations, Mr. McCurdy made statements to the effect that Mr. McCurdy was being directed by members who were aware of an available portfolio of intellectual property. It was public that Xockets' representative was affiliated with Xockets, and the Xockets portfolio was the only available portfolio that Xockets' representative was involved with at the time. Mr. McCurdy indicated he would go back to his members to consider next steps.

212. [redacted]

213. Given their AI-driven roles and the necessity of Xockets' technology to those roles as set forth above, NVIDIA and Microsoft constitute a large part of the demand for Xockets'

patented technology, and this market strength is exacerbated by their combination and agreement with and use of RPX, which counts among its members other companies that further make up the vast majority of the demand for Xockets' patented technology.

214. Since the time that RPX has become involved, Xockets has been prevented from obtaining a fair market price for its patents, which undeniably read on Microsoft's and NVIDIA's products, and neither Microsoft, nor NVIDIA, nor any of RPX's other members will negotiate at all with Xockets, thereby reducing output in the market for Xockets' patents to effectively zero.

VIII. ILLEGAL AGREEMENT BETWEEN THE DEFENDANTS

215. NVIDIA and Microsoft have agreed with each other and with RPX not to separately negotiate with Xockets and instead to only negotiate jointly via RPX. The individuals who entered into this agreement include but are not limited to those referenced in the preceding paragraphs. This buyers' cartel has allowed for price fixing by pushing the price and output below what would have been agreed to under normal market conditions. Indeed, the very purpose of RPX and the reason for joining RPX is to form groups of potential purchasers who can use collective purchasing power for technology inputs, something that NVIDIA and Microsoft understood and sought to employ via their unlawful agreement.

216. RPX's public statements, including but not limited to the statements referenced above, evidence that its platform is being employed for the purposes of a buyers' cartel. NVIDIA's and Microsoft's agreement to only negotiate as a buyers' cartel in the market for Xockets' patents has resulted in NVIDIA and Microsoft behaving in a manner contrary to their self-interest as horizontal competitors in that market. By negotiating individually, each would have had the opportunity to obtain a first mover advantage that in a functioning competitive market results in lower pricing and a competitive advantage against each other in the competition for Xockets' patents (including potentially even an exclusive license). This behavior also reflects an agreement

to either exercise or accumulate monopsony power and drive license fees and/or purchase costs substantially below market rates and/or to collectively refuse to license and/or purchase at all, which would drive Xockets out of business.

217. Xockets claims two separate bases for violation of the antitrust laws: (1) a conspiracy to restrain trade in violation of § 1 of the Sherman Act by all defendants, and (2) a conspiracy to monopolize (monopsonize) by all defendants in violation of § 2 of the Sherman Act. These antitrust claims are brought under §§ 4 and 16 of the Clayton Act (15 U.S.C. §§ 15 and 26): (a) to recover damages, including treble damages, sustained by Xockets as a result of its being injured in its business and property by reason of defendants' violations of the antitrust laws, particularly Sections 1 and 2 of the Sherman Act (15 U.S.C. §§ 1 and 2), (b) to obtain injunctive relief against threatened loss or damage as a result of such violations, and (c) to recover the expense of bringing and maintaining this action, including reasonable attorneys' fees.

COUNT I: VIOLATION OF SECTION 1 OF THE SHERMAN ACT BASED ON DEFENDANTS' CONSPIRACY IN RESTRAINT OF TRADE

218. Xockets incorporates by reference the preceding paragraphs as though fully set forth herein.

219. The facts set forth herein establish that a contract, combination, or conspiracy exists between and among at least RPX, NVIDIA, and Microsoft which restrained and continues to restrain trade or commerce among the several States in the market for the purchase, acquisition, or licensing of technology covered by Xockets' patents.

220. The agreement that NVIDIA and Microsoft act only through RPX to purchase, acquire, or license the Xockets patent portfolio forms the basis for the illegal contract, combination, or conspiracy. This agreement is to restrain output and fix prices below the market competitive price for Xockets' patented technology and/or to drive Xockets out of business.

221. An agreement exists between NVIDIA and Microsoft based on the alleged facts, including that NVIDIA and Microsoft directed and were therefore aware that RPX was negotiating on their behalf with Xockets and NVIDIA and Microsoft's respective refusal to discuss acquisition of the Xockets patent portfolio separately. The behavior of RPX evidencing that it is representing a buyers' cartel made up of at least NVIDIA and Microsoft also supports a reasonable inference that NVIDIA and Microsoft were acting in concert. RPX's public statements, discussed above, also support the existence of the buyers' cartel as they describe an invitation to concerted action, and to have RPX coordinate that concerted action.

222. The conspiracy, by virtue of Defendants' market power, is unreasonably restrictive of competition and Xockets suffered antitrust injury as a result.

223. The relevant product market is the market for purchase, acquisition, or licensing of technology covered by Xockets' patents.

224. Xockets' antitrust injury includes but is not limited to the fact that Defendants' conduct has destroyed the normal market forces that should have made it possible for Xockets to license or sell its technology. As a result of Defendants' conduct, Xockets has been unable to do so at normal market price, which has led to lost revenues and opportunities, and, if this conduct continues, Xockets will be driven out of business.

225. Moreover, Defendants' conduct not only harms competition with respect to the market for Xockets' patents, it also harms competition in the downstream markets for the market for GPU-enabled AI servers, which is controlled by NVIDIA, and the market for GPU-enabled AI platforms, which is controlled by Microsoft. As noted above, NVIDIA controls over 90% the market for GPU-enabled AI servers and Microsoft, through its partnership with Open AI, controls 70% of the market for GPU-enabled AI platforms. By driving down the costs of Xockets' patents,

Microsoft and NVIDIA can continue their dominance of these markets. If successful, this will harm invocation and allow NVIDIA and Microsoft to unilaterally increase prices within these markets.

226. Unless restrained by this Court, Defendants' unlawful conspiracy will continue to impose continuous injury and loss on Xockets' ability to sell or license its technology in a market free from such unlawful behavior.

COUNT II: VIOLATION OF SECTION 2 OF THE SHERMAN ACT BASED ON DEFENDANTS' CONSPIRACY TO CREATE OR MAINTAIN A MONOPSONY

227. Xockets incorporates by reference the preceding paragraphs as though fully set forth herein.

228. Defendants have combined or conspired in an attempt to obtain and/or in fact, have obtained monopsony power in the market for the purchase, acquisition, or licensing of technology covered by Xockets' patents, that power was or is being willfully acquired through Defendants' overt acts done with a specific intent to achieve monopsony power, and had an effect on a substantial amount of interstate commerce.

229. As detailed herein, the Defendants' anticompetitive conduct has eliminated competition between them for licenses to Xockets' patent portfolio with the effect of price fixing and more favorable non-monetary terms than possible in a market unaffected by such anticompetitive conduct.

230. Defendants either accumulated and are maintaining or are accumulating monopsony power over at least the relevant market and have used and are using that power to prevent Xockets from being able to sell or license its patent portfolio at normal market prices and to restrain quantities and fix prices at below the normal market rates.

231. Defendants' monopsony power is being or was willfully and intentionally acquired. The conspiracy amongst Defendants alleged herein was undertaken for the specific purpose of obtaining monopsony power over the market for the purchase, acquisition, or licensing of technology covered by Xockets' patents.

232. Defendants' monopsony power is not the result of good business skill or acumen, instead it is the product of their illegal conspiracy.

233. Unless restrained by this Court, Defendants' unlawful monopsonization of the market for the purchase, acquisition, or licensing of technology covered by Xockets' patents will continue to impose continuous injury and loss on Xockets' ability to sell or license its technology in a market free from such unlawful behavior. The impact being to illegally effect a substantial amount of interstate commerce.

234. Moreover, Defendants' conduct not only harms competition with respect to the market for Xockets' patents, it also harms competition in the downstream markets for the market for GPU-enabled AI servers, which is controlled by NVIDIA, and the market for GPU-enabled AI platforms, which is controlled by Microsoft. As noted above, NVIDIA controls over 90% the market for GPU-enabled AI servers and Microsoft, through its partnership with Open AI, controls 70% of the market for GPU-enabled AI platforms. By driving down the costs of Xockets' patents, Microsoft and NVIDIA can continue their dominance of these markets. If successful, this will harm invocation and allow NVIDIA and Microsoft to unilaterally increase prices within these markets.

DEMAND FOR JURY TRIAL

235. Pursuant to Rule 38 of the Federal Rules of Civil Procedure, Xockets demands a trial by jury on all issues so triable.

PRAYER FOR RELIEF

WHEREFORE, Xockets prays for judgment and requests that the Court find in its favor and against Defendants. Xockets respectfully requests that the Court enter preliminary and final orders, declarations, and judgments against Defendants as are necessary to provide Xockets with the following relief:

a. A judgment that Defendants' conduct alleged above is in violation of Sections 1 and 2 of the Sherman Act;

b. An award of threefold of the damages Plaintiff shall prove it has sustained on account of Defendants' violation of the antitrust laws, including, without limitation, pre-judgment and post-judgment interest;

c. The entry of an order preliminarily enjoining and restraining Defendants and their parents, affiliates, subsidiaries, officers, agents, servants, employees, attorneys, successors, and assigns and all those persons in active concert or participation with them or any of them, from violating the antitrust laws;

d. The entry of an order permanently enjoining and restraining Defendants and their parents, affiliates, subsidiaries, officers, agents, servants, employees, attorneys, successors, and assigns and all those persons in active concert or participation with them or any of them, from violating the antitrust laws; and

e. All further relief in law or in equity as the Court may deem just and proper.

Dated: September 5, 2024

Respectfully submitted,

/s/ Max Ciccarelli

Max Ciccarelli (SBN 00787242)
CICCARELLI LAW FIRM
100 N 6th Street, Suite 503
Waco, Texas 76701
Telephone: 214-444-8869
Email: max@ciccarellilawfirm.com

Jason G. Sheasby (*pro hac vice forthcoming*)
IRELL & MANELLA LLP
1800 Avenue of the Stars
Suite 900
Los Angeles, CA 90067
Tel.: 310.277.1010
Fax: 310.203.7199
Email: jsheasby@irell.com

Jamie H. McDole (SBN 24082049)
Lead Counsel
Phillip B. Philbin (SBN 15909020)
Michael D. Karson (SBN 24090198)
David W. Higer (SBN 24127850)
Miranda Y. Jones (SBN 24065519)
Grant Tucker (SBN 24121422)
Matthew L. Vitale (SBN 24137699)
WINSTEAD PC
2728 N. Harwood Street
Suite 500
Dallas, Texas 75201
Tel.: (214) 745-5400
Fax: (214) 745-5390
Email: jmcdole@winstead.com
pphilbin@winstead.com
mkarson@winstead.com
dhiger@winstead.com
mjones@winstead.com
gtucker@winstead.com
mvitale@winstead.com

Austin C. Teng (SBN 24093247)
Nadia E. Haghighatian (SBN 24087652)
WINSTEAD PC
600 W. 5th Street
Suite 900
Austin, Texas 78701
Tel.: (512) 370-2800
Fax: (512) 370-2850
Email: ateng@winstead.com
nhaghighatian@winstead.com

ATTORNEYS FOR PLAINTIFF XOCKETS, INC.